Reducing uncertainty in estimates of frequency distribution parameters using composite likelihood approach and copula-based bivariate distributions

Hemant Chowdhary[1] and Vijay P. Singh[2]

Received 7 August 2009; revised 16 July 2010; accepted 29 July 2010; published 11 November 2010.

[1] Conventional multivariate hydrological frequency analysis utilizes only the concurrent parts of data sets, leaving a lot of nonconcurrent data unutilized. Simultaneous inclusion of such nonconcurrent data can significantly reduce uncertainty in hydrologic design estimates. The methodology proposed in this paper allows varied length multivariate data to be combined and analyzed in an integrated framework through a "Composite Likelihood Approach." The method employs copula-based multivariate distributions in order to provide necessary flexibility of admitting arbitrary marginals. The paper presents the theoretical basis of the approach and highlights its advantages through two applications. A significant reduction in uncertainty in design flood quantiles of a relatively shorter flood series is achieved by utilizing an associated downstream flood data. The advantage of the methodology is further demonstrated by establishing significant information gain for six different combinations of Gaussian and non-Gaussian marginals. The proposed approach marks a paradigm shift in hydrologic design procedures, particularly for partially gauged basins, wherein a higher precision in hydrologic designs is achieved by leveraging associated information that has hitherto remained unutilized. It is opined that the approach will enable offsetting the impact of dwindling hydrological observation networks around the world by enhancing information that is derivable from existing networks.

Citation: Chowdhary, H., and V. P. Singh (2010), Reducing uncertainty in estimates of frequency distribution parameters using composite likelihood approach and copula-based bivariate distributions, *Water Resour. Res.*, *46*, W11516, doi:10.1029/2009WR008490.

1. Introduction

[2] Sufficiently long-term data is required for sound hydrological designs based on statistical analysis. Such long-term and good quality data is often a difficult proposition, especially in partially gauged basins around the world [*IAHS*, 2001]. Hydrological design estimates based on inadequate data lengths involve large uncertainties. As precision of hydrological estimates is proportional to data lengths, gauging networks need to be operated for sufficiently long periods, entailing cost consequences. There is, however, considerable potential for reducing uncertainties by extracting additional information from the associated hydrological data that are often available at or around the gauging station under consideration. As design methods in statistical hydrology have, to a large extent, been based on univariate analyses, such additional information has not been harnessed for the purpose of uncertainty reduction. A few nonstructural measures, such as station-year method for precipitation [*Buishand*, 1984] or index flood method for flood frequency [*Dalrymple*, 1960], and regional regression techniques have been evolved for reducing biases and uncertainties by pooling regional information. These methods, excepting those developed later, e.g., the generalized least squares technique [*Stedinger and Tasker*, 1985], do not fully incorporate the dependence characteristics inherent in such associated data.

[3] Furthermore, hydrological information invariably contains a lot of staggered data of which a bulk of nonconcurrent data remains unutilized. There is a huge untapped potential for reducing parameter uncertainty by considering a multivariate framework that allows for simultaneous consideration of partially concurrent information in an integrated manner. A few studies in the past have utilized such partially concurrent or incomplete bivariate or multivariate data sets. These studies, however, employed conventional bivariate or multivariate frequency distributions that are restrictive in having to choose the marginals from the same distribution types, limiting the usage of the methodology. These studies have been for normal distribution using bivariate or trivariate normal distributions [*Lord*, 1955; *Edgett*, 1956; *Anderson*, 1957; *Fiering*, 1962; *Matalas and Jacobs*, 1964; *Rueda*, 1981], for gamma distribution using bivariate gamma distribution [*Clarke*, 1980], for largest extreme value distribution using different forms of bivariate and trivariate Gumbel distributions [*Rueda*, 1981; *Raynal-Villasenor*, 1985; *Escalante*

[1]Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, Louisiana, USA.
[2]Department of Biological and Agricultural Engineering and Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas, USA.

0043-1397/10/2009WR008490

and Raynal-Villasenor, 1998], for Weibull distribution using bivariate Weibull and mixed Weibull distributions [*Escalante*, 2007], or for the general extreme value distribution using bivariate and trivariate general extreme value distributions [*Escalante and Raynal-Villasenor*, 2008; *Raynal-Villasenor and Salas* 2008]. Applications made in these studies were, e.g., for extending a shorter annual streamflow record using a longer precipitation record, extending a shorter flood record by using longer flood records from one or more adjoining stations, or for regional flood frequency analysis.

[4] In all of the above studies, the bivariate or multivariate distributions comprised the same types of marginals, such as normal, largest extreme value (Gumbel), general extreme value, Weibull, or mixed Weibull. Intuitively, the choice of marginals must be dictated by the type of distribution that best represents the data under consideration. It is likely that arbitrarily different marginals constitute the multivariate process of interest. The copula-based distributions can advantageously provide means for combining such marginals. The methodology presented here is motivated by this flexibility offered by copula models, with the objective of developing a copula-based "Composite Likelihood Approach" for reducing uncertainty in the estimates of frequency distribution parameters. The approach comprises univariate, bivariate, and/or multivariate likelihood components as per periods of concurrency in the available data. This paper presents the theoretical basis of the proposed approach, leading to expressions for information gain that is accruable by integrating the nonconcurrent associated hydrological information that has hitherto remained untapped. The two applications included in the paper illustrate the usefulness of the approach. In the first application, a significant reduction in uncertainty of design flood quantiles for a relatively shorter flood data series is attained by simultaneously utilizing a longer flood data series from another downstream station. The second application involves quantifying the "expected information gain" for six different combinations of Gaussian and non-Gaussian marginals and conventional and copula-based bivariate distributions. The paper is organized in four sections. Following the introductory remarks and a brief review, the objectives are laid out in section 1. Section 2 provides necessary details of the composite likelihood approach, including assumptions and applicability, and steps for estimating information gain. The advantages of the approach are demonstrated by presenting two applications in sections 3 and 4. The important gains achievable by this approach are discussed in section 5, and conclusions and future research directions are summarized in section 6. It is opined that this proposed approach will provide a framework for improving the precision of hydrologic design estimates, particularly those based on inadequate data, at virtually no extra cost, as it utilizes existing associated hydrological data that has hitherto remained unharnessed for this purpose.

2. Composite Likelihood Approach

[5] The maximum likelihood estimation method is frequently applied in hydrologic applications owing to its large sample properties of, in general, yielding consistent estimates with minimum variance. Under certain regularity conditions, these estimates for large samples are considered as good as any other estimates [*Mood et al.*, 1974]. The asymptotic variances can be obtained from the Cramer-Rao theorem that provides the lower bound on the dispersion of parameter estimates. Estimates for small samples are also invariably approximated on this basis and have found general acceptance in practice. These estimates are assumed to be normally distributed, with estimated values as the mean vector and asymptotic variance-covariance matrix representing dispersion. The proposed composite likelihood approach considers both the concurrent and nonconcurrent parts of an associated multivariate data set in an integrated manner and provides more precise parameter estimates. In a way, this approach provides a mechanism to transfer information from an associated data series to a relatively shorter data series under consideration. This information gain leads to a reduction in uncertainty and can be quantified as the ratio of reciprocals of variances of estimates resulting from composite and simple likelihood approaches.

[6] The methodology assumes homogeneity, serial independence, and stationary properties among individual variables and in their dependence characteristics. It would be applicable to most hydrological designs, as they are also invariably based on similar assumptions. Starting with a simple likelihood function, the composite likelihood approach is outlined next, providing quantitative expressions for information gain.

2.1. Dispersion Matrix Based on Simple Likelihood Function

[7] Considering a multivariate random variable $\mathbf{X} = (X_1, X_2, \ldots, X_k)$, having its k-dimensional probability density function (pdf) $f(\mathbf{x}|\boldsymbol{\psi})$ in R_k real-space and r-dimensional parameter vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_r)$, the likelihood function L with respect to n independent and identically distributed (iid) observations is given as

$$L = L(\boldsymbol{\psi}|\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\psi}).$$

The corresponding log likelihood function l is given as

$$l = \log L(\boldsymbol{\psi}|\mathbf{x}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i|\boldsymbol{\psi}). \quad (1)$$

For brevity, $f(\mathbf{x}|\boldsymbol{\psi})$ is hereafter written as $f(\mathbf{x})$, $L(\boldsymbol{\psi}|\mathbf{x})$ as $L(\mathbf{x})$ or L, and $\log L(\boldsymbol{\psi}|\mathbf{x})$ as $\log L(\mathbf{x})$ or $\log L$ or l. The maximum likelihood estimates $\hat{\psi}_p$ for $p = 1{:}r$ are obtained by maximizing the log likelihood function by solving the system of equations given by

$$\frac{\partial \log L}{\partial \psi_p} = \sum_{i=1}^{n} \frac{\partial \log f(\mathbf{x}_i)}{\partial \psi_p} = 0.$$

The Cramer-Rao Lower Bound (CRLB) provides the lower bound on the dispersion of parameters obtained by any estimation method. Such variance-covariance matrix is inversely proportional to the Fisher information matrix. The (p,q)th element of the Fisher information matrix with respect to the above multivariate likelihood function L, as adapted from

Wilks [1962], *Rao* [1973], and *Cox and Hinkley* [1974] and elaborated by *Chowdhary* [2010], is given by

$$i^{p,q} = \mathrm{E}\left[\frac{\partial \log L(\mathbf{X})}{\partial \psi_p}\frac{\partial \log L(\mathbf{X})}{\partial \psi_q}\right] = n\mathrm{E}\left[\frac{\partial \log f(\mathbf{X})}{\partial \psi_p}\frac{\partial \log f(\mathbf{X})}{\partial \psi_q}\right] = n\mathrm{E}\big[S_p(\mathbf{X})\,S_q(\mathbf{X})\big] = n\,\mathrm{E}\left[-\frac{\partial^2 \log f(\mathbf{X})}{\partial \psi_p \partial \psi_q}\right] = n\mathrm{E}\big[-S_{pq}(\mathbf{X})\big], \tag{2}$$

where $S_p(\mathbf{x}) = \frac{\partial \log f(\mathbf{x})}{\partial \psi_p}$ for $p = 1{:}r$ is the score function and $S_{pq}(\mathbf{x}) = \frac{\partial^2 \log f(\mathbf{x})}{\partial \psi_p \partial \psi_q}$ is its first derivative with respect to parameter $q = 1{:}r$.

[8] The Fisher information matrix $\mathbf{I}$ is then obtained as

$$\mathbf{I} = \| \, i^{p,q} \, \|_{rxr} = n \| \, a^{p,q} \, \| = n\,\mathbf{A},$$

where $\mathbf{A} = \|a^{p,q}\|_{rxr}$ and $a^{p,q}$, the information content derivable from a single observation of $\mathbf{X}$, using equation (2), is given by

$$a^{p,q} = \frac{1}{n} i^{p,q} = \mathrm{E}\big[S_p(\mathbf{X})\,S_q(\mathbf{X})\big] = \mathrm{E}\big[-S_{pq}(\mathbf{X})\big]. \tag{3}$$

The variance-covariance matrix $\mathbf{VC} = \|vc^{pq}\|_{rxr}$ for the asymptotically efficient estimates is then obtained by inverting the Fisher information matrix as

$$\mathbf{VC} = \mathbf{I}^{-1} = \frac{1}{n}\mathbf{A}^{-1} = \frac{1}{n}\mathbf{B}, \tag{4}$$

with matrix $\mathbf{B} = \|b^{p,q}\| = (\mathbf{A})^{-1}$.

2.1.1. Special Case I: Univariate Distribution

[9] As a special case, considering $\mathbf{X} = X$ to be a univariate random variable, the elements of the information matrix, using equation (2), are given as

$$i_X^{p,q} = \mathrm{E}\left[\frac{\partial \log L_X(X)}{\partial \psi_p}\frac{\partial \log L_X(X)}{\partial \psi_q}\right] = n\mathrm{E}\left[\frac{\partial \log f(X)}{\partial \psi_p}\frac{\partial \log f(X)}{\partial \psi_q}\right] = n\mathrm{E}\big[S_p(X)S_q(X)\big] = n\mathrm{E}\left[-\frac{\partial^2 \log f(X)}{\partial \psi_p \partial \psi_q}\right] = n\mathrm{E}\big[-S_{pq}(X)\big]. \tag{5a}$$

The expectation terms in the above equalities are obtained by either algebraic or numerical integration as

$$\mathrm{E}\big[S_p(X)S_q(X)\big] = \int_{-\infty}^{+\infty} S_p(x)S_q(x)f(x)\,\mathrm{d}x \quad \text{or} \quad \mathrm{E}\big[S_{pq}(X)\big] = \int_{-\infty}^{+\infty} S_{pq}(x)f(x)\mathrm{d}x. \tag{5b}$$

The resulting Fisher information matrix $\mathbf{I}_X$ is given by

$$\mathbf{I}_X = \| \, i_X^{p,q} \, \|_{rxr} = n \| \, a_X^{p,q} \, \| = n\mathbf{A}_X,$$

where $\mathbf{A}_X = \|a_X^{p,q}\|_{rxr}$ and again $a_X^{p,q}$, the information content derivable from a single observation of X, using equation (5a) is given by

$$a_X^{p,q} = \frac{1}{n} i_X^{p,q} = \mathrm{E}\big[S_p(X)S_q(X)\big] = \mathrm{E}\big[-S_{pq}(X)\big]. \tag{6}$$

For asymptotically efficient estimates, the corresponding variance-covariance matrix $\mathbf{VC}_X = \|vc_X^{pq}\|_{rxr}$, taking a matrix $\mathbf{B}_X = \|b_X^{p,q}\| = (\mathbf{A}_X)^{-1}$, is given as

$$\mathbf{VC}_X = (\mathbf{I}_X)^{-1} = \frac{1}{n}(\mathbf{A}_X)^{-1} = \frac{1}{n}\mathbf{B}_X. \tag{7}$$

2.1.2. Special Case II: Bivariate Distribution

[10] Considering $\mathbf{X} = (X, Y)$ to be a bivariate random variable, the elements of the information matrix, using equation (2), are given as

$$i_{XY}^{p,q} = \mathrm{E}\left[\frac{\partial \log L_{XY}(X,Y)}{\partial \psi_p}\frac{\partial \log L_{XY}(X,Y)}{\partial \psi_q}\right] = n\mathrm{E}\left[\frac{\partial \log f(X,Y)}{\partial \psi_p}\frac{\partial \log f(X,Y)}{\partial \psi_q}\right] = n\mathrm{E}\big[S_p(X,Y)S_q(X,Y)\big] = n\mathrm{E}\left[-\frac{\partial^2 \log f(X,Y)}{\partial \psi_p \partial \psi_q}\right] = n\mathrm{E}\big[-S_{pq}(X,Y)\big]. \tag{8a}$$

The expectation terms in the above equalities are obtained by either algebraic or numerical integration as

$$\mathrm{E}\big[S_p(X,Y)S_q(X,Y)\big] = \int_{-\infty}^{+\infty} S_p(x,y)S_q(x,y)f(x,y)\mathrm{d}x\,\mathrm{d}y \quad \text{or} \quad \mathrm{E}\big[S_{pq}(X,Y)\big] = \int_{-\infty}^{+\infty} S_{pq}(x,y)f(x,y)\mathrm{d}x\,\mathrm{d}y. \tag{8b}$$

The resulting Fisher information matrix is given by

$$\mathbf{I}_{XY} = \| \, i_{XY}^{p,q} \, \|_{rxr} = n \| \, a_{XY}^{p,q} \, \| = n\mathbf{A}_{XY},$$

where $\mathbf{A}_{XY} = \|a_{XY}^{p,q}\|_{rxr}$ and again $a_{XY}^{p,q}$, the information content derivable from a single observation of (X,Y), using equation (8a) is given by

$$a_{XY}^{p,q} = \frac{1}{n} i_{XY}^{p,q} = \mathrm{E}\big[S_p(X,Y)S_q(X,Y)\big] = \mathrm{E}\big[-S_{pq}(X,Y)\big]. \tag{9}$$

For asymptotically efficient estimates, the corresponding variance-covariance matrix $\mathbf{VC}_{XY} = \|vc_{XY}^{pq}\|_{rxr}$, considering a matrix $\mathbf{B}_{XY} = \|b_{XY}^{p,q}\| = (\mathbf{A}_{XY})^{-1}$, is given as

$$\mathbf{VC}_{XY} = (\mathbf{I}_{XY})^{-1} = \frac{1}{n}(\mathbf{A}_{XY})^{-1} = \frac{1}{n}\mathbf{B}_{XY}. \tag{10}$$

For the sake of brevity, subscript XY has been used in lieu of X,Y in the above equations.

2.2. Composite Likelihood Function

[11] The likelihood of a composite event, comprising some concurrent and some exclusive (nonconcurrent) periods of X and Y, as shown in Figure 1, is called here the composite likelihood function. N_X and N_Y are the total available lengths (sample sizes) of the two data series individually. Of these lengths, n_{XY} is the concurrent period, and n_X and n_Y are the exclusive periods. These periods have been shown contiguous for the sake of clarity only and without loss of generality these represent cases of intermittent data availability as well. Let $f_X(x;\boldsymbol{\delta})$ and $f_Y(y;\boldsymbol{\eta})$ represent marginal pdfs and

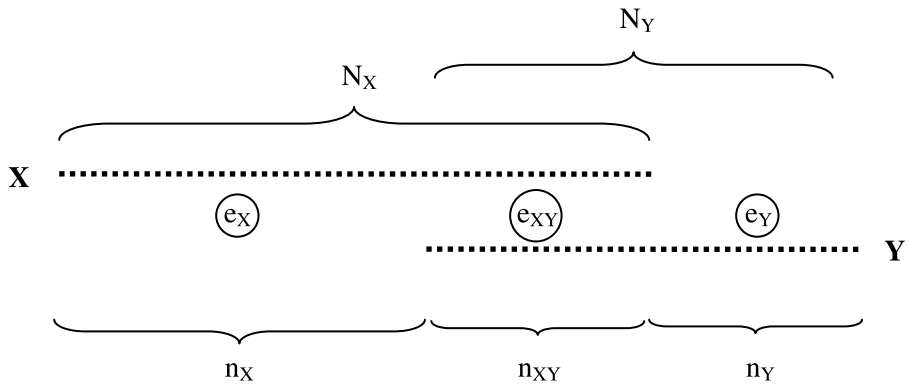

Figure 1. Arrangement of an incomplete sample of bivariate random variable (X,Y).

$f_{X,Y}(x,y;\boldsymbol{\psi})$ be the bivariate pdf of (X,Y). Here $\boldsymbol{\delta} = \{\delta_1,\delta_2,\ldots\delta_{r_\delta}\}$, $\boldsymbol{\eta} = \{\eta_1, \eta_2,\ldots\eta_{r_\eta}\}$ and $\boldsymbol{\psi} = \{\boldsymbol{\delta},\boldsymbol{\eta},\boldsymbol{\theta}\} = \{\psi_1,\psi_2,\ldots\psi_r\}$ are the parameter vectors of these distributions, respectively. $\boldsymbol{\theta} = \{\theta_1,\theta_2,\ldots\theta_{r_\theta}\}$ is the association parameter vector appearing in the bivariate pdf. For purposes of brevity, these pdfs are hereafter written as $f(x)$, $f(y)$, and $f(x,y)$.

[12] The whole realization in Figure 1 can be considered a composite event e_C, comprising three independent events, e_X, e_Y, and e_{XY}, that occur in the two exclusive and one concurrent period, respectively. Let the likelihood functions corresponding to all observations of these three events be L_X, L_Y, and L_{XY}, and their joint likelihood, termed here as the "composite likelihood function," be L_C. Owing to the mutual independence of these three events, the composite likelihood can be expressed as the product of their individual likelihoods [*Rao*, 1973] as

$$L_C = L_X L_Y L_{XY}.$$

The corresponding log likelihood is the sum of the individual log likelihoods as

$$l_C = \log L_C = \log L_X + \log L_Y + \log L_{XY}. \tag{11a}$$

This can easily be extended by induction to multivariate cases. Taking, e.g., a trivariate random variable (X,Y,Z), the composite log likelihood, using similar notations, can be expressed as

$$l_C = \log L_C = \log L_X + \log L_Y + \log L_Z + \log L_{XY} + \log L_{YZ} + \log L_{ZX} + \log L_{XYZ}.$$

Depending on the presence of certain exclusive and concurrent periods in the trivariate data set, the respective components in the above equality can be retained for further computations.

2.3. Dispersion Matrix Based on Composite Likelihood Function

[13] The maximum likelihood estimates $\hat{\psi}_p$ for $p = 1{:}r$ in case of composite events, such as in Figure 1, are obtained by maximizing the composite log likelihood function and solving the system of equations given by

$$\frac{\partial \log L_C}{\partial \psi_p} = \frac{\partial(\log L_X + \log L_Y + \log L_{XY})}{\partial \psi_p} = 0. \tag{11b}$$

As in equation (2), the (p,q)th element of the Fisher information matrix, with respect to the composite likelihood function L_C, is given by

$$i_C^{p,q} = \mathrm{E}\left[\frac{\partial \log L_C}{\partial \psi_p}\frac{\partial \log L_C}{\partial \psi_q}\right].$$

Using equation (11a), this can be simplified as

$$\begin{aligned} i_C^{p,q} &= \mathrm{E}\left[\frac{\partial(\log L_X + \log L_Y + \log L_{XY})}{\partial \psi_p}\frac{\partial(\log L_X + \log L_Y + \log L_{XY})}{\partial \psi_q}\right] \\ &= \mathrm{E}\left\{\sum_{i\in\{X,Y,XY\}}\left[\frac{\partial \log L_i}{\partial \psi_p}\frac{\partial \log L_i}{\partial \psi_q}\right] + \sum_{i,j\in\{X,Y,XY\};i\neq j}\left[\frac{\partial \log L_i}{\partial \psi_p}\frac{\partial \log L_j}{\partial \psi_q}\right]\right\} \\ &= \sum_{i\in\{X,Y,XY\}}\mathrm{E}\left[\frac{\partial \log L_i}{\partial \psi_p}\frac{\partial \log L_i}{\partial \psi_q}\right] + \sum_{i,j\in\{X,Y,XY\};i\neq j}\left[\mathrm{E}\left(\frac{\partial \log L_i}{\partial \psi_p}\right)\mathrm{E}\left(\frac{\partial \log L_j}{\partial \psi_q}\right)\right]. \end{aligned}$$

In the above equality, the expectation of the product within the second summation is expressed as the product of the expectations, owing to the independence among events e_i and e_j, where $e_i, e_j \in \{e_X, e_Y, e_{XY}\}$ and $e_i \neq e_j$. Further, simple algebraic steps [*Chowdhary*, 2010] show that all these expectations are equal to zero and thus

$$i_C^{p,q} = \mathrm{E}\left[\frac{\partial \log L_X}{\partial \psi_p}\frac{\partial \log L_X}{\partial \psi_q}\right] + \mathrm{E}\left[\frac{\partial \log L_Y}{\partial \psi_p}\frac{\partial \log L_Y}{\partial \psi_q}\right] + \mathrm{E}\left[\frac{\partial \log L_{XY}}{\partial \psi_p}\frac{\partial \log L_{XY}}{\partial \psi_q}\right].$$

The three expectation terms in the above equality correspond to the elements of the Fisher information matrix for univariate and bivariate distributions, as given in equation (5a) and (8a), and thus

$$i_C^{p,q} = i_X^{p,q} + i_Y^{p,q} + i_{XY}^{p,q}.$$

This summation of information from the constituent events of a composite event is in agreement with the Fisher information

of independent observations (here observations are from e_X, e_Y, and e_{XY} events) being additive [*Rao*, 1973]. Considering the length of the concurrent period $n_{XY} > 0$, the above result can be expressed in terms of $a^{p,q}$s, using equations (6) and (9), as

$$i_C^{p,q} = n_X a_X^{p,q} + n_Y a_Y^{p,q} + n_{XY} a_{XY}^{p,q} = n_{XY}\left[\frac{n_X}{n_{XY}} a_X{}^{p,q} + \frac{n_Y}{n_{XY}} a_Y^{p,q} + a_{XY}^{p,q}\right]$$
$$= n_{XY}\left[\left(\frac{N_X}{n_{XY}} - 1\right) a_X^{p,q} + \left(\frac{N_Y}{n_{XY}} - 1\right) a_Y^{p,q} + a_{XY}^{p,q}\right].$$

Taking the ratios of total lengths of X and Y and the concurrent period as $m_X = N_X/n_{XY}$ and $m_Y = N_Y/n_{XY}$, respectively, the elements of the information matrix for composite events becomes

$$i_C^{p,q} = n_{XY}[(m_X - 1) a_X^{p,q} + (m_Y - 1) a_Y^{p,q} + a_{XY}^{p,q}]. \quad (12)$$

The resulting Fisher information matrix $\mathbf{I}_C$ is given by

$$\mathbf{I}_C = \| i_C^{p,q} \|_{rxr} = n_{XY} \| a_C^{p,q} \| = n_{XY} \mathbf{A}_C,$$

where $\mathbf{A}_C = \|a_C^{p,q}\|_{rxr}$ and $a_C^{p,q}$, the information content corresponding to a single concurrent bivariate observation and proportional contributions from the exclusive univariate portions is given as

$$a_C^{p,q} = \frac{1}{n_{XY}} i_C^{p,q} = (m_X - 1) a_X^{p,q} + (m_Y - 1) a_Y^{p,q} + a_{XY}^{p,q}. \quad (13)$$

For the asymptotically efficient estimates, the corresponding variance-covariance matrix $\mathbf{VC}_C = \|vc_C^{pq}\|_{rxr}$, taking a matrix $\mathbf{B}_C = \|b_C^{p,q}\| = (\mathbf{A}_C)^{-1}$, is given as

$$\mathbf{VC}_C = (\mathbf{I}_C)^{-1} = \frac{1}{n_{XY}} (\mathbf{A}_C)^{-1} = \frac{1}{n_{XY}} \mathbf{B}_C. \quad (14)$$

The results in equation (12) are easily extendable by induction to multivariate cases. Taking, e.g., a trivariate random variable (X,Y,Z), the elements of the Fisher information matrix that are the basis of all further computations, using similar notations as above, can be expressed as

$$i_C^{p,q} = n_X a_X^{p,q} + n_Y a_Y^{p,q} + n_Z a_Z^{p,q} + n_{XY} a_{XY}^{p,q} + n_{YZ} a_{YZ}^{p,q} + n_{ZX} a_{ZX}^{p,q} + n_{XYZ} a_{XYZ}^{p,q}$$
$$= n_{XYZ}\begin{bmatrix} (m_X - m_{XY} - m_{XZ} + 1) a_X^{p,q} + (m_Y - m_{XY} - m_{YZ} + 1) a_Y^{p,q} \\ + (m_Z - m_{ZX} - m_{YZ} + 1) a_Z^{p,q} + (m_{XY} - 1) a_{XY}^{p,q} \\ + (m_{YZ} - 1) a_{YZ}^{p,q} + (m_{ZX} - 1) a_{ZX}^{p,q} + a_{XYZ}^{p,q} \end{bmatrix} \quad (15)$$

Various length ratios used in the above equality are intuitive and can be obtained by considering the total univariate and bivariate lengths as N_X, N_Y, N_Z and N_{XY}, N_{YZ}, N_{ZX}, respectively, and the trivariate concurrent length as n_{XYZ} and taking $m_X = N_X/n_{XYZ}$, $m_Y = N_Y/n_{XYZ}$, $m_Z = N_Z/n_{XYZ}$, $m_{XY} = N_{XY}/n_{XYZ}$, $m_{YZ} = N_{YZ}/n_{XYZ}$, and $m_{ZX} = N_{ZX}/n_{XYZ}$. If only certain concurrent and exclusive periods are present, then the corresponding components only need to be retained for further computations. The above formulation assumes the presence of a trivariate concurrent period, i.e., $n_{XYZ} > 0$ to be true. The available trivariate data set can still be advantageously pooled for reducing uncertainty even if the trivariate concurrent period is absent. In such case, the first equality of equation (15), in which n_{XYZ} is not factored out, can be used after dropping the trivariate term.

2.4. Modeling Concurrent Periods Using Copulas

[14] The joint distribution function $f(x,y)$ appearing in equation (8b) is required for computing the information content derivable from a concurrent event e_{XY} that is one of the constituent of a composite event e_C. As mentioned earlier in section 1, information gain for parameter estimates of a few distributions has been investigated in the past. One of the limitations in all those studies has been in having the marginals from the same type of distribution for modeling the concurrent bivariate or trivariate periods. Application potential can increase tremendously if this limitation is overcome. The copula-based distributions have an advantage over conventional distributions with respect to admitting arbitrary marginals and can therefore be advantageously employed for representing concurrent periods of the composite events.

[15] The joint probability in copula-based distributions is expressed in terms of marginal probabilities and more advantageously in terms of uniform marginals. This theory has been in vogue for some time, especially with respect to actuarial science and finance applications. In recent years, copula-based distributions are being increasingly employed in the field of hydrologic engineering. Several illustrative and review studies [*Favre et al.*, 2004; *Salvadori and De Michele*, 2004; *De Michele et al.*, 2005; *Grimaldi et al.*, 2005; *Zhang and Singh*, 2006, 2007; *Genest and Favre*, 2007; *Poulin et al.*, 2007; *Salvadori and De Michele*, 2007; *Serinaldi and Grimaldi*, 2007; *Kao and Govindaraju*, 2007a, 2007b, 2008; *Serinaldi*, 2009; *Samaniego et al.* 2010; *Vandenberghe et al.*, 2010; among others] provide elaborate discussion on copula applications related to flow and rainfall variables. Reference may be made to *Joe* [1997] and *Nelsen* [2006] for theoretical details on dependence and copulas.

[16] Considering a bivariate random variable (X,Y), its joint cumulative distribution function (cdf) $F(x,y)$, in terms of probability transformed or standard uniform variates $u = F_X(x)$ and $v = F_Y(y)$, is given as

$$F(x,y) = C_\theta[F_X(x), F_Y(y)] = C_\theta(u,v) \quad (16)$$

where $F_X(x)$ and $F_Y(y)$ are marginal cdfs and $C_\theta : [0,1]\times[0,1] \rightarrow [0,1]$, a mapping function, is the "copula" that combines marginal probabilities into joint probability. In turn, it means that a valid joint distribution model for (X,Y) is obtained whenever the three constituents $(C, F_X$, and $F_Y)$ are chosen from given parametric families, viz.,

$$F_X(x;\boldsymbol{\delta}), \quad F_Y(y;\boldsymbol{\eta}), \quad C_\theta(u,v;\boldsymbol{\theta}),$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are the parameter vectors of marginal distributions, and $\boldsymbol{\theta}$ is the dependence parameter vector.

[17] By double differentiating equation (16), the copula-based joint pdf, involving copula density $c_\theta(u,v)$ and marginal pdfs $f(x)$ and $f(y)$, is obtained as

$$f(x,y) = f(x) f(y) c_\theta(u,v). \quad (17)$$

There are several copula classes and families, such as Archimedean, meta-elliptic, extreme value, and miscellaneous class. A number of models under each of these categories

provide a great deal of flexibility in choosing copula models that may be suitable for any particular application. The functional forms of joint cdfs and pdfs of various copula models along with their dependence characteristics and parameter spaces are obtainable from any standard text on copula, such as *Joe* [1997] and *Nelsen* [2006]. Selection of copula models is discussed in one of the applications presented later in section 3.

2.5. Information Gain

[18] Simplified expressions for the information gain for asymptotically efficient estimates based on the composite likelihood approach can be derived in terms of elements of variance-covariance matrices. As information is a reciprocal measure of dispersion, the ratio of reciprocals of the corresponding elements of variance-covariance matrices obtained from the composite likelihood and univariate approaches would signify information gain. Terming such element-wise ratio corresponding to estimates of pth and qth parameters of the distribution function for X as relative information RI_X^{pq}, it can be expressed in terms of the elements of variance-covariance matrices of univariate and composite events given in equations (7) and (14) as

$$RI_X^{pq} = \frac{\frac{1}{vc_C^{pq}}}{\frac{1}{vc_X^{pq}}} = \frac{vc_X^{pq}}{vc_C^{pq}} = \frac{\frac{1}{N_X} b_X^{pq}}{\frac{1}{n_{XY}} b_C^{pq}} = \frac{1}{\frac{N_X}{n_{XY}}} \left(\frac{b_X^{pq}}{b_C^{pq}} \right) = \frac{1}{m_X} \left(\frac{b_X^{pq}}{b_C^{pq}} \right). \quad (18)$$

Similarly, the relative information RI_Y^{pq}, corresponding to asymptotically efficient estimates of distribution function for variable Y, is given as

$$RI_Y^{pq} = \frac{\frac{1}{vc_C^{pq}}}{\frac{1}{vc_Y^{pq}}} = \frac{vc_Y^{pq}}{vc_C^{pq}} = \frac{\frac{1}{N_Y} b_Y^{pq}}{\frac{1}{n_{XY}} b_C^{pq}} = \frac{1}{\frac{N_Y}{n_{XY}}} \left(\frac{b_Y^{pq}}{b_C^{pq}} \right) = \frac{1}{m_Y} \left(\frac{b_Y^{pq}}{b_C^{pq}} \right). \quad (19)$$

When sample estimates of variance terms b_X^{pq}, b_Y^{pq}, and b_C^{pq} are used in equations (18) and (19) for a data set, the resulting relative information pertains to that particular sample. On the other hand, "expected information gain" is obtained if expected variances are used in the above equalities. The expected information gain is an important result, as it indicates the extent of additional information that is accruable on an average for any given combination of marginals and copula model. The second application in section 4 illustrates such advantage for six different cases. The relative information in excess of unity indicates additional information yielded by the composite likelihood approach, resulting in a reduction of uncertainty in parameter estimates. In all cases presented in this article, the relative information is equal to or greater than unity and significantly more than unity in most cases. This demonstrates the main advantage of this proposed approach that systematically integrates available hydrologic data and provides more precise parameters and design estimates.

[19] Another way of looking at relative information gains is in terms of virtual data augmentation. Variances being inversely proportional to data lengths, a reduction in variance is equivalent to augmenting the data by certain length. Assuming the effective (or virtual) lengths for X or Y to be N_X^C and N_Y^C, these relate to the information gains as

$$RI_X^{pq} = \frac{vc_X^{pq}}{vc_C^{pq}} = \frac{\frac{1}{N_X} b_X^{pq}}{\frac{1}{N_X^C} b_X^{pq}} = \frac{N_X^C}{N_X} \text{ and } RI_Y^{pq} = \frac{vc_Y^{pq}}{vc_C^{pq}} = \frac{\frac{1}{N_Y} b_Y^{pq}}{\frac{1}{N_Y^C} b_Y^{pq}} = \frac{N_Y^C}{N_Y}.$$

In other words, the augmented lengths of the two series are equal to the product of relative information and the actual available lengths. Thus, the percent information gain indirectly indicates an equivalent gain in the length of the data series.

3. Application I: Flood Quantiles Based on Composite Likelihood Approach

[20] This application illustrates the prime advantage of the composite likelihood approach by obtaining more precise design flood quantiles. The uncertainty in flood quantiles for a relatively shorter upstream river gauging station is significantly reduced by utilizing associated data from a longer annual flow series available from a downstream station. After describing the data set and identifying potential marginals, a copula model is selected on the basis of three parameter estimation methods and several graphical and analytical inference procedures, including assessing overall and tail dependence characteristics. The composite likelihood approach is then employed to show significant reduction in uncertainty in flood quantiles for the shorter series.

3.1. Data Set and Potential Marginals

[21] The annual peak flow data for two USGS river gauging stations, Alderson (03183500) and Buckeye (03182500), on Greenbrier River in West Virginia State considered in this application are obtained from the online National Water Information System of the USGS (http://nwis.waterdata.usgs.gov). The Greenbrier River is a tributary of New River in southeastern part of the state and is approximately 265 km long. Flowing into New, Kanawha, and Ohio rivers, it is part of the Mississippi River watershed. The Alderson station located at 37°43′27″ latitude and −80°38′30″ longitude is 122 km downstream of the Buckeye station which is located at 38°11′09″ latitude and −80°07′51″ longitude. These stations command 3500 and 1400 km^2 of drainage areas and have riverbeds at about 466 and 636 m above sea level, respectively. Annual peak flows for Alderson station for 112 years from 1896 to 2007 and for Buckeye station for 78 years from 1930 to 2007 are available. These time series at Alderson ($X = Q_A$) and Buckeye ($Y = Q_B$) stations are shown in Figure 2 (left).

[22] Several candidate distributions such as normal (NOR), two- and three-parameter lognormal (LN2 and LN3), two-parameter gamma (G2), Pearson type III (P3), log-Pearson type III (LP3), largest extreme value (LEV), and two- and three-parameter Weibull (W2 and W3) are first considered for fitting the two data series on a univariate basis. LEV and LP3 distributions are found to adequately represent annual peak flows at Alderson and Buckeye stations, respectively, on the basis of Kolmogorov-Smirnov, Anderson Darling, and Chi-square test statistics and the overall fit of observed and computed quantiles (QQ plots). The overlay of probability

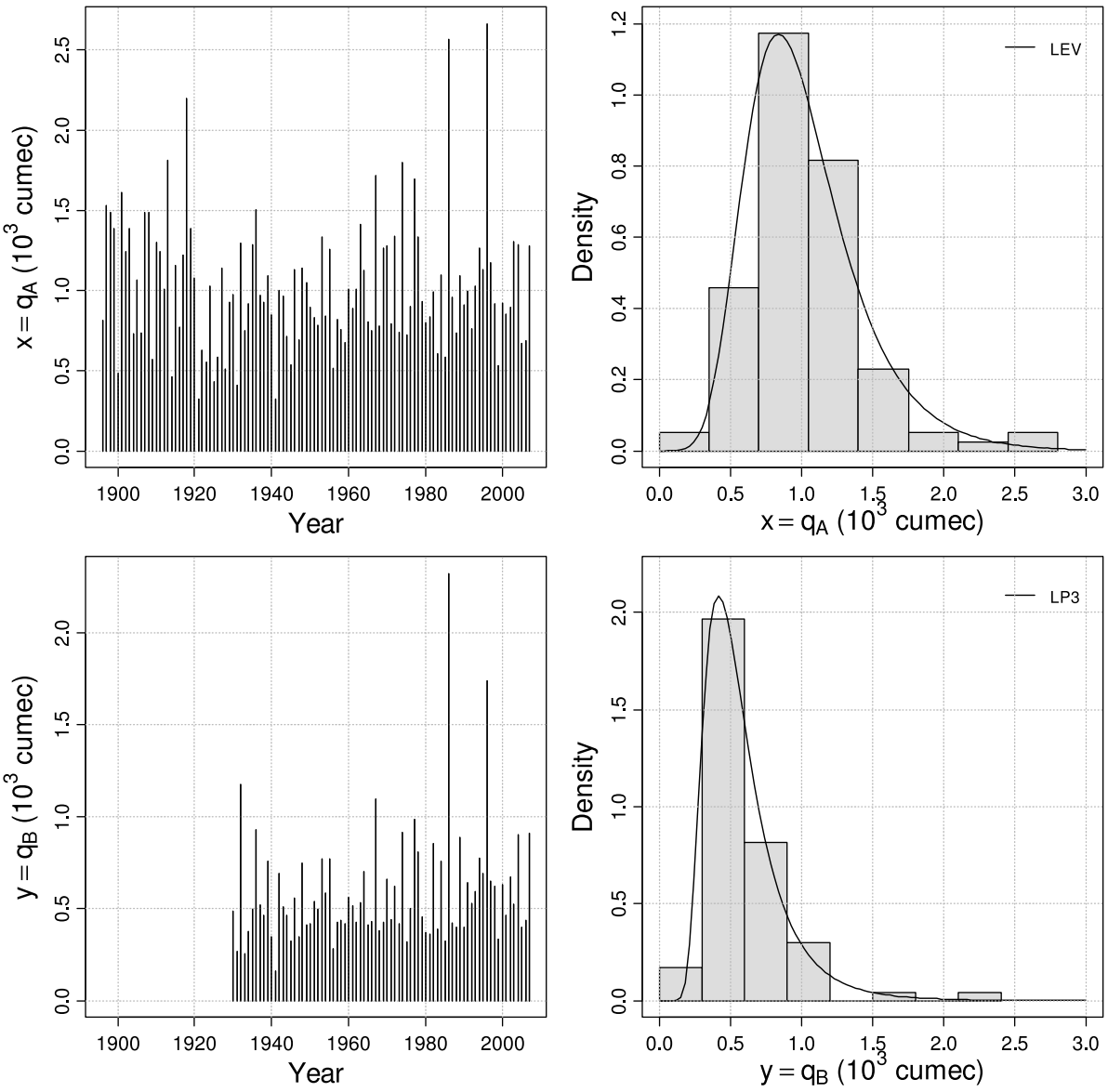

Figure 2. (left) Time series and (right) histograms of annual peak flows at Alderson ($X = Q_A$) and Buckeye ($Y = Q_B$) stations. The probability density functions of the fitted LEV and LP3 distributions are plotted over the respective histograms.

density functions of these distributions and the corresponding histograms is shown in Figure 2 (right).

3.2. Bivariate Dependence and Copula Models

[23] The concurrent period of 78 years among the two annual peak flow series is considered for obtaining copula models representing dependence and joint distribution of the bivariate data. The scatterplots of annual peak flows at Alderson and Buckeye stations, of their scaled ranks (scaled by n + 1), and of their computed probabilities $\hat{F}_X$ and $\hat{F}_Y$ (based on the fitted LEV and LP3 distributions), along with respective histograms are shown in Figure 3. As expected for the annual peak flows from two stations on the same river, all these three scatterplots indicate a strong positive dependence. The sample estimates of the association measures, the Pearson correlation coefficient ρ, the Kendall tau τ, and the Spearman rho ρ_s, 0.887, 0.643, and 0.823, respectively, with corresponding p values $\ll$1.0e-20, $\ll$1.0e-20, and 2.15e-20, corroborate this assertion. Further qualitative assessment of dependence from Chi plot and K plot as proposed by *Fisher and Switzer* [2001] and *Genest and Boies* [2003], respectively, reaffirms the significant positive dependence. Lower tail independence and upper tail dependence is evident by considering bottom left and top right quadrants of these plots exclusively, as suggested by *Abberger* [2005].

[24] Archimedean copulas such as Ali-Mikhail-Haq (AMH), Clayton or Cook-Johnson, Frank, extreme value copulas as Gumble-Hougaard (GH) and Galambos (GH being an Archimedean copula also) and miscellaneous copula as Farlie-Gumbel-Morgenstern (FGM) are considered for modeling the joint probability of the two annual peak flow data. On the basis of the Kendall tau value of the sample, 0.643, and the similarity with lower and upper tail dependence features, the two extreme value copulas, Galambos and GH, make

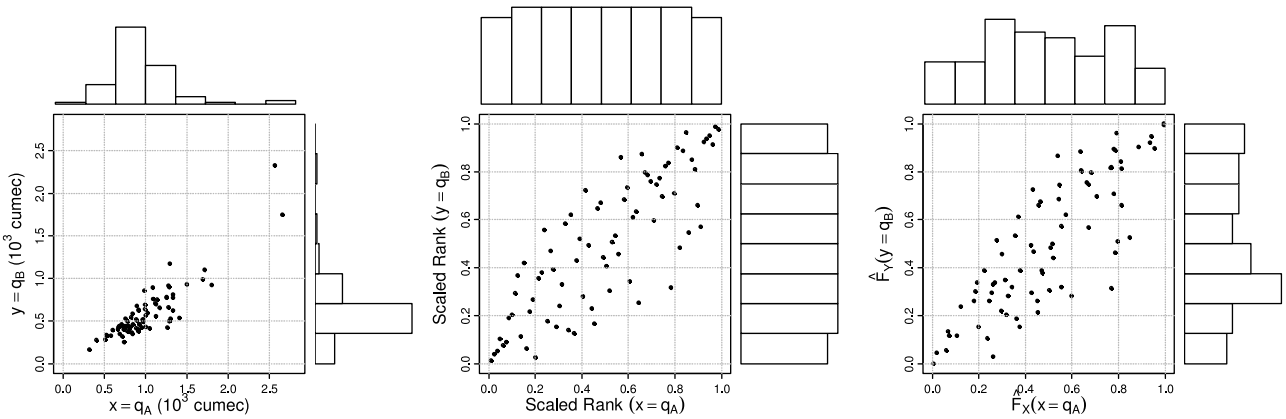

Figure 3. Scatter plots and histograms of the bivariate annual peak flows at Alderson ($X = Q_A$) and Buckeye ($Y = Q_B$) stations in (a) original domain, (b) as scaled ranks, and (c) as LEV and LP3 computed probabilities.

plausible copula models. In order to appreciate the relative suitability of Clayton and Frank copulas, these are also short-listed owing to their comprehensive dependence characteristics. The AMH and FGM copulas are excluded from further consideration as these admit Kendall tau in ranges of −0.1817 to 0.3333, and −0.222 to 0.222, respectively, both of which fall significantly short of covering the sample Kendall tau estimate.

[25] The dependence parameters for copula families under consideration are estimated by three methods: (1) moment-like method of inversion of association measures (MOM), (2) maximum pseudolikelihood (MPL), and (3) "inference from margins" (IFM) [*Genest and Favre*, 2007]. For the MOM method, the estimates of dependence parameter θ can be obtained from its relationships with Kendall tau τ or Spearman rho ρ_s. The MPL and IFM method-based estimates are obtained by maximizing the likelihood function, involving empirical and computed marginal probabilities, respectively. These estimates, the respective standard errors and the maximized log likelihood values (LL_{max}) are given in Table 1. The sample Kendall tau used in MOM method and those obtained by inverting estimated dependence parameter θ in MPL and IFM methods are also given in Table 1. It may be seen that standard errors from the MPL method are much lower than those for the MOM method. The errors are only slightly larger for the IFM method as compared to the MPL method, except for the Clayton copula for which it is much larger. The Galambos and GH copulas have very similar maximized log likelihood values that are much higher than other copulas. Thus, from the point of view of likelihood values and standard errors, the Galambos and GH copulas obtained by the MPL method may be preferable.

3.3. Assessment of Copula Fitting

[26] The relative suitability of various copula models is ascertained in multiple ways by employing (1) graphical methods, (2) error statistics, and (3) analytical goodness-of-fit tests.

3.3.1. Graphical Methods

[27] First, the scatterplot of observed bivariate data is compared by superimposing it on sets of large number of random bivariate samples generated on the basis of various hypothesized copulas. Sets of 500 samples are generated for each of the four copula families, utilizing MOM, MPL, and IFM method-based parameters. Figure 4 shows one such representative comparison for the MPL method, and it is apparent that the spread of observed and generated data matches well for the Galambos and GH copulas, whereas for the Clayton and Frank copulas the match is poor, particularly for higher peak flows. Second, comparison of empirical probabilities and computed probabilities revealed the extent of fitting of copula surface with the empirical prob-

Table 1. Copula Dependence Parameter Estimates and Error Statistics for Annual Peak Flows at Alderson and Buckeye Stations

Method/Copula Family	Theta ($\hat{\theta}$)	Tau ($\hat{\tau}$)	Standard Error	LL_{max}	RMSE	MN-A-ERR	MX-A-ERR
			MOM				
Clayton	3.655	0.643	0.693	–	0.023	0.018	0.054
Frank	9.314		1.451	–	0.016	0.013	0.043
Galambos	2.115		0.345	–	0.013	0.012	0.025
GH	2.828		0.346	–	0.013	0.012	0.025
			MPL				
Clayton	2.468	0.552	0.274	40.263	0.032	0.025	0.064
Frank	8.789	0.630	0.159	42.957	0.017	0.014	0.044
Galambos	2.062	0.639	0.193	48.447	0.014	0.012	0.026
GH	2.769	0.639	0.172	48.283	0.014	0.012	0.026
			IFM				
Clayton	2.650	0.570	8.447	38.462	0.030	0.024	0.062
Frank	9.977	0.665	0.174	46.088	0.016	0.012	0.042
Galambos	2.218	0.659	0.290	51.979	0.012	0.011	0.023
GH	2.928	0.658	0.252	51.825	0.012	0.011	0.023

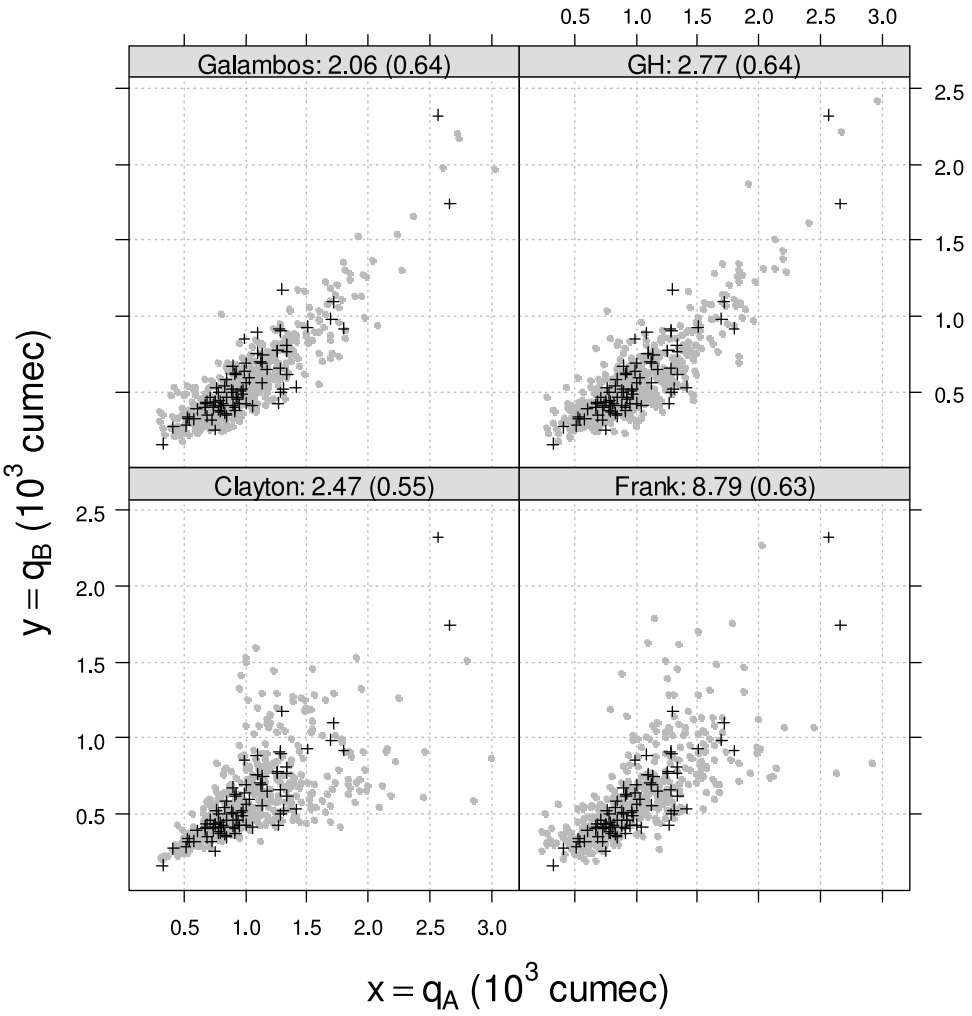

Figure 4. Comparison of observed annual peak flows at Alderson and Buckeye stations and the MPL method-based random samples for various copulas. Solid circles are random samples (size 500), and plus symbols represent observed data. Numbers in name strips are dependence parameter estimates with corresponding Kendall tau values in parenthesis.

ability surface. The matching for both Galambos and GH copulas was very good in every case, whereas differences progressively increased for the Frank and Clayton copulas, particularly for larger joint probabilities. Third, a graphical comparison of empirical and computed probability distributions, $K_n(w)$ and $K_{\theta_n}(w)$, of the bivariate probability integral transform (BIPIT) variate $W = C(U,V)$, and that of observed and expected order statistics of W in the form of generalized K plots as proposed by *Genest and Favre* [2007] is done. All these plots showed very good matching for the Galambos and GH copulas, followed by progressively inferior matching for the Frank and Clayton copulas. For brevity, the nature of these comparisons is represented by K plots for the MPL method in Figure 5.

3.3.2. Error Statistics

[28] Comparison of error statistics, such as root mean square error (RMSE), mean absolute error (MN-A-ERR), and maximum absolute error (MX-A-ERR) given in Table 1 also provides a measure of fit of copula models to the joint empirical probability. All these statistics for the three methods indicate better suitability of the Galambos and GH copulas as against progressively larger errors for the Frank and Clayton copulas. It may be observed that errors for the three methods are comparable, except for the Clayton copula for which the MPL and IFM methods yield much larger errors, owing to lower dependence returned by these methods.

3.3.3. Analytical Goodness-of-Fit Tests

[29] Three goodness-of-fit test statistics have been employed to formally test the adequacy of the hypothesized copulas. The first, comparing empirical and parametric copula probabilities based on the process $\sqrt{n}(C_n - C_{\theta_n})$, is the Cramer-von Mises type statistic $\mathcal{CM}_n$ proposed by *Fermanian* [2005]. The other two, providing comparison of the empirical and theoretical probabilities of W on the basis of the process $K_n(w) = \sqrt{n}\{K_n(w) - K_{\theta_n}(w)\}$, are the Cramer-von Mises and Kolmogorov type statistics $\mathcal{S}_n$ and $\mathcal{T}_n$ given by *Genest et al.* [2006] as variants of those proposed by *Wang and Wells* [2000]. Tests based on these three statistics, involving a parametric bootstrap procedure, are conducted for the three parameter estimation methods. For this, 10,000 samples of the size of the observed bivariate peak flow data set (i.e., 78) are simulated, except for the Galambos copula for which only 1000 sets are generated owing to a larger computational time requirement. The results of three such simulation runs

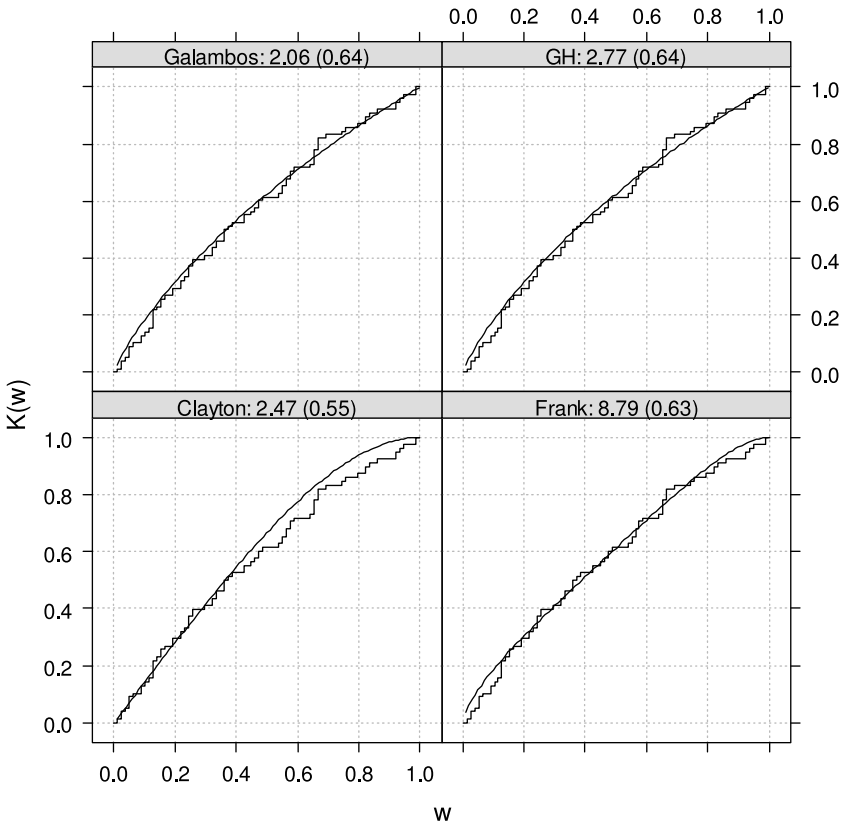

Figure 5. Graphical goodness-of-fit tests for various copulas using K plots for the MPL method-based estimates for annual peak flows at Alderson and Buckeye stations. Step functions are the empirical distributions K_n (w), and curves are the theoretical distributions K_{θ_n} (w) of the bivariate integral transform variable $W = C(U, V)$. Numbers in name strips are dependence parameter estimates with corresponding Kendall tau values in parenthesis.

for only the MPL method are given in Table 2 for the sake of brevity. Very low p values returned in each instance for the Clayton copula provide overwhelming evidence for its rejection at a 5% significance level. On the other hand, no evidence is found for rejection of the Galambos, GH, and Frank copulas. The very high p values for the Galambos and GH copulas suggest that these may be preferred over the Frank copula.

[30] All the graphical and analytical goodness-of-fit test results thus suggest Galambos and GH copulas as suitable models for the bivariate data under consideration. Although either of these copulas would make a viable model for this data set, the GH copula is finally selected, owing to its frequent use in hydrological applications. And in view of smaller standard error (Table 1), the MPL estimation method is employed for estimating the parameters.

3.4. Composite Likelihood-Based Flood Quantiles

[31] Considering LEV and LP3 marginals and the GH copula as selected above, the distribution parameter estimates

Table 2. Goodness-of-Fit Statistics for Hypothesized Copulas With Respect to the MPL Method[a]

	$\mathcal{CM}_n$			$\mathcal{S}_n$			$\mathcal{T}_n$		
Copula	Obs. Stat.	S^*	P	Obs. Stat.	S^*	P	Obs. Stat.	S^*	P
		4.693	**0.002**		0.207	**0.029**		1.075	**0.031**
Clayton	8.499	4.648	**0.001**	0.233	0.207	**0.029**	1.132	1.075	**0.028**
		4.785	**0.001**		0.209	**0.030**		1.073	**0.028**
		2.551	0.321		0.122	0.416		0.902	0.494
Frank	1.764	2.542	0.317	0.080	0.121	0.413	0.844	0.901	0.506
		2.561	0.316		0.122	0.414		0.902	0.500
		2.872	0.824		0.125	0.403		0.870	0.570
Galambos	1.124	2.920	0.826	0.076	0.132	0.418	0.739	0.886	0.602
		2.847	0.818		0.126	0.393		0.914	0.587
		2.812	0.820		0.125	0.408		0.878	0.598
GH	1.134	2.771	0.812	0.077	0.124	0.392	0.739	0.880	0.587
		2.796	0.804		0.124	0.392		0.882	0.583

[a] S^* implies critical value of the test statistic at 5% significance level, and P indicates the p values of the observed test statistic (Obs. Stat.).

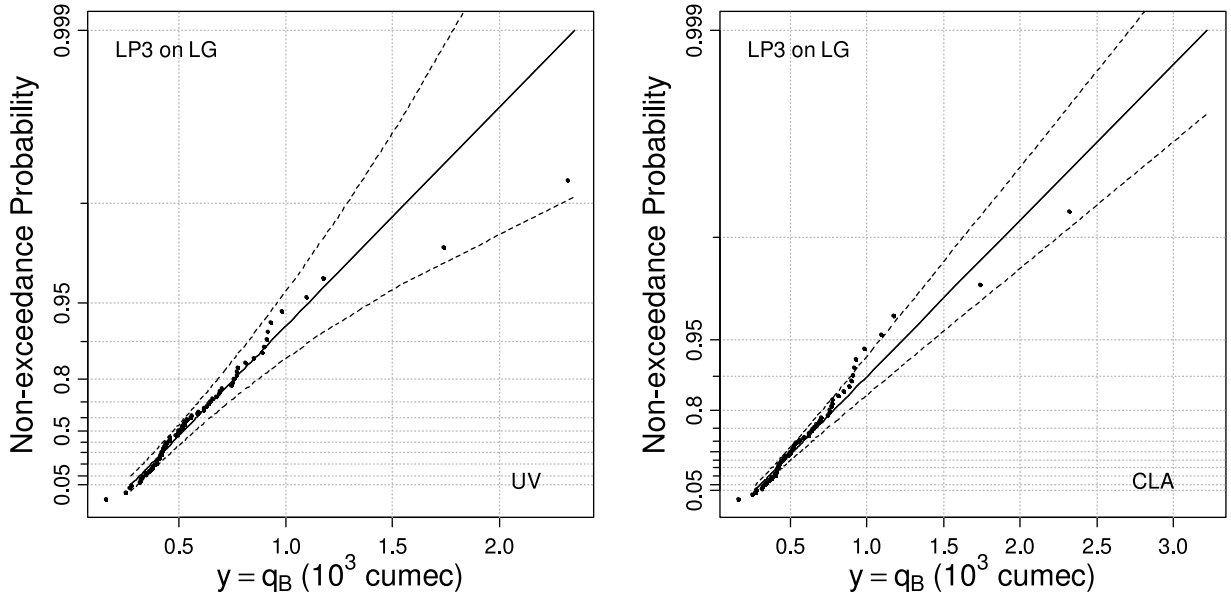

Figure 6. Comparison of probability plots for annual peak flows at Buckeye station obtained on the basis of purely univariate analysis (annotated as "UV") and by composite likelihood approach (annotated as "CLA").

and associated variance-covariance matrix based on the composite likelihood approach are obtained from equations (11b) and (14), respectively. The interval estimates for 100, 200, and 500 year return period flood quantiles are obtained on the basis of these parameters and dispersion characteristics. These quantiles along with associated standard errors and confidence widths corresponding to a coverage probability of 95% are given in Table 3. The quantile estimates, standard errors, and confidence widths obtained purely on a univariate basis in section 3.1 are also given in Table 3. About 30% reduction in standard errors and confidence widths for the three quantiles shows significant benefits that are accruable upon considering both annual peak flow series in a composite fashion as compared to the usual univariate analysis. Comparison of LP3 probability plots on a log-gamma probability paper resulting from these two approaches is shown in Figure 6. The narrower confidence band in the case of composite likelihood approach graphically illustrates the benefit of the approach. This application pertained to a moderate length ratio of $m_X = 112/78 = 1.43$. The percent uncertainty reduction is expected to be even higher for greater length ratios of 2–3 that are typically expected for an actual design application.

[32] Apart from reduction in uncertainty in terms of standard errors and confidence widths, a few more error statistics, such as root mean square error (RMSE), standard error of fit (SEF), bias (BIAS), and mean absolute relative deviations (MARD), are also computed for the two approaches. These error estimates along with percent changes are given in Table 4. Again, a reduction of about 10%–20% in all these error estimates is achieved by employing the composite likelihood approach.

4. Application II: Quantifying Expected Information Gain

[33] This application highlights the benefits of the approach by showing the extent of the "expected information gain" that is accruable for certain combinations of marginals and a copula-based bivariate distribution. No observed data is involved in this application, as it deals with expected information gains that are obtainable from equations (5b) and (8b). The application is illustrated by considering a simplified arrangement of the composite event shown in Figure 1 wherein only the exclusive period of X and the concurrent period of (X,Y) are present, i.e., for $n_Y = 0$ and $N_Y = n_{XY}$. The variable X, having both exclusive and concurrent periods, would hereafter be called "longer series," and Y, having only the concurrent period, would be called "shorter series." The objective here is to see if the uncertainty in distribution parameter estimates for X and Y decreases by employing the composite likelihood approach. Six cases involving different

Table 3. Comparison of Uncertainty in Flood Quantiles for Buckeye Station Obtained From Purely Univariate Approach (UV) and Composite Likelihood Approach (CLA)

	Computed Quantiles and Error Statistics (10^3 cumec)								
	UV Based			CLA Based			% Change		
Return Period (years)	Quantile	Standard Error	Confidence Width	Quantile	Standard Error	Confidence Width	Quantile	Standard Error	Confidence Width
100	1.58	0.20	0.80	1.72	0.15	0.58	−8.4	27.8	27.6
200	1.81	0.26	1.04	2.01	0.19	0.74	−10.9	29.0	29.5
500	2.14	0.36	1.44	2.45	0.25	0.99	−14.6	30.7	31.6

Table 4. Comparison of Error Statistics for LP3 Distribution Fitting for Buckeye Station Obtained for Purely Univariate Approach (UV) and Composite Likelihood Approach (CLA)

Error Statistic	Error Estimates		% Change
	UV Based	CLA Based	
RMSE	2.94	2.31	21.6
SEF	2.98	2.34	21.6
BIAS	−0.15	−0.12	19.6
MARD	4.58	4.11	10.2

marginals and bivariate distributions, as listed in Table 5, are analyzed and compared for the gains by shorter and longer series. The four marginals considered are normal (NOR), largest extreme value or Gumbel (LEV), gamma (G2), and log-Pearson Type III (LP3). Cases I and II involve conventional bivariate normal and Gumbel Type A [*Gumbel and Mustafi*, 1967] distributions, respectively. These two cases have been studied earlier as well [*Rueda*, 1981; *Raynal-Villasenor*, 1985] but are presented here for purposes of comparison and completeness. Admitting different marginals, Cases III to VI employ the Frank copula for modeling concurrent bivariate periods. The selection of specific univariate and bivariate distributions here, including the Frank copula, is solely for illustrative purposes and without loss of generality, results for the expected information gain are obtainable for any other marginal distributions and copula types.

[34] The combinations of marginals considered in these six cases have direct relevance to hydrologic applications. For example, the combination of two NOR distributions (Case I) may be suitable for normally distributed variables, such as annual flow volumes and (or) annual rainfall from adjoining gauging stations or watersheds. The two LEV distributions (Case II) may be of use when dealing with extreme value variables, such as annual peak flood and (or) annual peak storm rainfall, or maximal water quality parameters that exhibit a certain level of association. The combination of NOR and LEV or G2 marginals (Cases III and IV) may be useful where data for an extreme value variable, such as annual peak flood, is of shorter length but an associated normally distributed variable, such as annual precipitation or annual flow volume, is available for a longer period. Similarly, the combination of G2 and LEV distributions (Case V) or LEV and LP3 distributions (Case VI) could be employed where data for an extreme value variable such as annual peak flood is of shorter length but another associated extreme value distributed annual peak flood data from nearby station on the same or adjoining river is available for a longer period. Such possibility of arbitrarily combining different marginals can be advantageous for improving precision of estimates in many other hydrologic applications, involving precipitation, evaporation, flow, soil moisture, groundwater, and/or water quality variables, where one or more variables have limited data availability but there is sufficiently long-term availability of one or more other associated variables from same or nearby observation stations.

[35] The expressions of pdf and cdf for the four marginals and bivariate normal and Gumbel Type A distributions can be obtained from any standard text on distribution functions. The mean and variance for NOR distribution are denoted as μ_i and σ_i^2, location and scale parameters for LEV distribution as γ_i and α_i, and scale and shape parameters for G2 distribution as α_i and β_i, for $i \in \{X,Y\}$, and location, scale, and shape parameters for the LP3 distribution for Y as γ_Y, α_Y, and β_Y, respectively. The joint cdf for the Frank copula-based bivariate distribution is obtained from equation (16) wherein $C(u,v)$ is taken from any standard text on copulas as indicated in section 2.4. The corresponding joint pdfs can be obtained from equation (17) after obtaining copula density $c_\theta(u,v)$ by double differentiating $C(u,v)$.

4.1. Asymptotic Variances

[36] Expressions for exact variances can be derived for parameter estimates that can be expressed in explicit form. Obtaining expressions for parameter estimates that cannot be expressed in closed form may become a formidable task. Asymptotic variances are used as a good approximation in such situations when sample size is large enough. *Rueda* [1981] showed that even for sample sizes as small as 10 and 20, asymptotic variances provide acceptable estimates of exact variances for bivariate normal and bivariate Gumbel Type A distributions. Asymptotic variances can be obtained for a univariate or a bivariate distribution as the inverse of the information matrix. Elements of the information matrix can be obtained from either the third or the fifth equality of equations (5a) and (8a) while computing expectation terms through integration as in equations (5b) and (8b).

4.1.1. Information Matrices for Univariate Distributions

[37] Considering n observations of a univariate random variable X, the elements of information matrices for normal, largest extreme value, gamma and log-Pearson Type III distributions can be obtained from equation (5a) after a few algebraic steps and knowing the moment generating functions of the distributions. As information matrices are symmetric, it suffices to write the elements of upper diagonal matrices. For estimates of mean and variance for normal distribution, the information matrix is given as

$$I_{NOR}(\mu,\sigma^2) = n\begin{Vmatrix} \frac{1}{\sigma^2} & 0 \\ & \frac{1}{2(\sigma^2)^2} \end{Vmatrix}. \tag{20}$$

For the largest extreme value distribution, the information matrix for location and scale parameters is given as

$$I_{LEV}(\gamma,\alpha) = \frac{n}{\alpha^2}\begin{Vmatrix} 1 & -0.4228 \\ & 1.8237 \end{Vmatrix}. \tag{21}$$

Table 5. Six Cases for Which Application of Composite Likelihood Approach Has Been Illustrated

Case	Marginal Distribution of		Bivariate Distribution
	Longer Series X	Shorter Series Y	
I	NOR	NOR	Bivariate normal
II	LEV	LEV	Bivariate Gumbel Type A
III	NOR	LEV	Frank copula
IV	NOR	G2	Frank copula
V	G2	LEV	Frank copula
VI	LEV	LP3	Frank copula

For the gamma distribution, the information matrix for scale and shape parameters is given as

$$I_{G2}(\alpha,\beta) = n\begin{Vmatrix} \frac{\beta}{\alpha^2} & \frac{1}{\alpha} \\ & \psi'(\beta) \end{Vmatrix}. \quad (22)$$

Here $\psi'(\beta)$ is called "psi" or "trigamma" function, given by

$$\psi'(\beta) = \frac{\partial \psi(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\left[\frac{\Gamma'(\beta)}{\Gamma(\beta)}\right].$$

The information matrix for location, scale, and shape parameters for the log-Pearson Type III distribution is given as

$$I_{LP3}(\gamma,\alpha,\beta) = n\begin{Vmatrix} \frac{1}{\alpha^2(\beta-2)} & \frac{1}{\alpha^2} & \frac{1}{\alpha^2(\beta-1)} \\ & \frac{\beta}{\alpha^2} & \frac{1}{\alpha} \\ & & \psi'(\beta) \end{Vmatrix}. \quad (23)$$

4.1.2. Information Matrices for Bivariate Distributions

[38] Obtaining elements of information matrices for bivariate distributions quickly becomes an involved process as one moves from bivariate normal to other non-Gaussian distributions, such as the bivariate Gumbel Type A or the copula-based distributions. For the bivariate normal distribution, these elements of information matrix can be easily derived in closed form. Considering n observations of a normally distributed bivariate random variable (X,Y), the information matrix with respect to means, variances and dependence parameter is given as

$$I_{XY}\left(\mu_X,\sigma_X^2,\mu_Y,\sigma_y^2,\theta\right) = n\begin{Vmatrix} \frac{1}{\sigma_X^2(1-\theta^2)} & 0 & \frac{-\theta}{\sigma_X\sigma_Y(1-\theta^2)} & 0 & 0 \\ & \frac{(2-\theta^2)}{4(\sigma_X^2)^2(1-\theta^2)} & 0 & \frac{-\theta^2}{4\sigma_X^2\sigma_Y^2(1-\theta^2)} & \frac{-\theta}{2\sigma_X^2(1-\theta^2)} \\ & & \frac{1}{\sigma_Y^2(1-\theta^2)} & 0 & 0 \\ & & & \frac{(2-\theta^2)}{4(\sigma_Y^2)^2(1-\theta^2)} & \frac{-\theta}{2\sigma_Y^2(1-\theta^2)} \\ & & & & \frac{1+\theta^2}{(1-\theta^2)^2} \end{Vmatrix}. \quad (24)$$

For the bivariate Gumbel Type A distribution, getting elements of information matrix in closed form becomes cumbersome and therefore expectation terms in equation (8a) are obtained through numerical integration. It is easier, in this case, to work with the product of score functions $S_p(x,y)$ and $S_q(x,y)$ rather than their derivatives $S_{pq}(x,y)$. The score functions for the bivariate Gumbel Type A distribution, with respect to location, scale, and association parameters $\psi_p = \{\gamma_X,\alpha_X,\gamma_Y,\alpha_Y,\theta\}$ for $p = 1{:}5$, are given as

$$S_p(x,y) = \frac{\partial \log f(x,y)}{\partial \psi_p} = \frac{\partial}{\partial \psi_p}\left[-\log\alpha_X - \log\alpha_Y + \log f_{Z_X,Z_Y}(z_X,z_Y)\right],$$

where $Z_X = \frac{X-\gamma_X}{\alpha_X}$ and $Z_Y = \frac{X-\gamma_Y}{\alpha_Y}$ are the standard LEV variates and $F_{Z_X,Z_Y}(z_X,z_Y)$ is the standard bivariate Gumbel Type A pdf obtained from its cdf,

$$F_{Z_X,Z_Y}(z_X,z_Y) = \exp\left[-e^{-z_X} - e^{-z_Y} + \theta(-e^{z_X} + -e^{z_Y})^{-1}\right].$$

Similarly, for copula-based distributions also, the expectations in equation (8a) are difficult to obtain algebraically and recourse is taken to numerical integration. Using equation (17), the derivatives $S_p(x,y)$ for copula-based distributions can be obtained as

$$\begin{aligned} S_p(x,y) &= \frac{\partial \log f(x,y)}{\partial \psi_p} = \frac{\partial}{\partial \psi_p}\left[\log f(x) + \log f(y) + \log c_\theta(u,v)\right] \\ &= \frac{\partial \log f(x)}{\partial \psi_p} + \frac{\partial \log f(y)}{\partial \psi_p} + \frac{\partial \log c_\theta(u,v)}{\partial \psi_p}. \end{aligned}$$

The first two derivatives in the last equality are the score functions of the two marginals, and therefore when $\psi_p \in \{\boldsymbol{\delta}\}$ or $\psi_p \in \{\boldsymbol{\eta}\}$ the score function for the other marginal becomes zero. After obtaining score functions $S_p(x,y)$ with respect to all the parameters, the elements of information matrix can be obtained numerically using equation (8a).

4.2. Expected Information Gain

[39] The relative information for parameter estimates of marginals in all the six cases is computed on the basis of equations (18) and (19), involving length ratios $m_X = N_X/n_{XY}$ and $m_Y = N_Y/n_{XY}$ and dependence parameter θ. For the simplified arrangement in Figure 1 under consideration, $n_Y = 0$ and $N_Y = n_{XY}$ results in $m_Y = 1$. The results can be generalized for any individual lengths of longer and shorter series such that $m_X \geq 1$. The results are therefore presented as charts of relative information plotted against the length ratio m_X and for various levels of association. As the significance of the dependence parameter θ as a measure of association is different for different distributions, it makes it an unsuitable reference while presenting or comparing results of information gain. In order to assess the role of dependence, it is desirable to relate the information gain with certain association measures that have common meaning for different distributions. The parametric Pearson correlation coefficient ρ or rank-based nonparametric measures, such as Spearman correlation coefficient ρ_s and Kendall τ, could be used for the purpose. Using the correspondence between θ and the association measures ρ, ρ_s, and τ as given by *Nelsen* [2006],

Table 6. Percent Information Gain for Distribution Parameter Estimates for Longer Series X

	Marginals		% Information Gain for Parameter Estimates of X at $m_X = 1$							
			For Moderate ρ				For High ρ			
Case	X	Y	ρ	$\hat{\mu}_X/\hat{\gamma}_Y$	$\hat{\sigma}_X^2/\hat{\alpha}_X$	$\hat{\beta}_X$	ρ	$\hat{\mu}_X/\hat{\gamma}_Y$	$\hat{\sigma}_X^2/\hat{\alpha}_X$	$\hat{\beta}_X$
I	**NOR**	NOR	0.7	0	0	–	0.9	0	0	–
II	**LEV**	LEV	0.6	2	12	–	0.67	3	29	–
III	**NOR**	LEV	0.7	18	8	–	0.9	65	41	–
IV	**NOR**	G2	0.7	16	6	–	0.9	51	24	–
V	**G2**	LEV	0.7	–	8	8	0.9	–	30	29
VI	**LEV**	LP3	0.4	18	12	–	0.5	40	27	–

their interrelationships are established with respect to the six cases under consideration. The values of ρ and equivalent ρ_s, in parenthesis, shown in all the plots for reference and use have been obtained on the basis of these relationships. The Pearson correlation coefficient ρ is, however, not preferred as it can be misleading at times for the non-Gaussian distributed data. Nevertheless, since ρ is a more familiar measure to many practitioners, the same is retained in all the plots for an easy reference.

[40] For Cases I–III, the relative information is a function of length ratio m_X and association level only. The results in these cases are therefore valid for any member of the involved marginals, i.e., for any values of the constituent parameters. For Cases IV–VI, the relative information is also a function of shape parameters of the gamma and log-Pearson Type III distributions. Therefore, computations for these cases may be done for specific shape parameter that may be of interest. For illustration, the results for shape parameter $\beta_Y = 12$ are presented here. Furthermore, for Case VI, the relationship between ρ and θ also involves the scale parameter of LP3 distribution and a value of $\alpha_Y = 0.25$ has been used. These specific values of scale and shape parameters have been taken for illustrative purposes only and without loss of generality, the expected information gain for any other admissible values can be obtained.

[41] The expected information gain for all the cases has been computed for m_X ranging from 1 to 5. Such a range would easily cover most practical hydrological applications, e.g., when a shorter series is available for only about 20 years and a longer series is available for about 40–60 years. Further, the association levels in increments of 0.1 of the Pearson correlation coefficient ρ are considered in the range of 0.1–0.9, wherever admissible. A common characteristic observed in the results of all the cases is that the relative information for parameters of longer series decreases with the increase in values of m_X, whereas for shorter series it increases with increasing m_X. This is expected as there is relatively lesser proportional contribution from the shorter to longer series when the value of m_X increases. On the other hand, a shorter series gains more from relatively larger contribution of longer series when the value of m_X increases. For purposes of comparison and reporting, a value of $m_X = 1$ for the longer series is used, indicating the scenario of maximal gain for such marginals if another marginal having about the same length is available. For the shorter series, a value of $m_X = 3$ is used as it is more likely to expect values about this in real-life situations. Similarly, $\rho = 0.7$ (i.e., coefficient of determination $R^2 = 0.49$), a moderate value to expect for the supposedly associated variables under consideration, is used as a practical reference value. However, a higher dependence is expected among related variables and for that reason results are also reported and discussed with reference to a higher value of $\rho = 0.9$ ($R^2 = 0.81$) wherever admissible. The information gain for longer and shorter series is summarized in Tables 6 and 7, respectively. The results for the six cases are discussed in the following six subsections.

4.2.1. Case I: Normal Marginals

[42] The objective in this case is to see if there is a gain in information for a normally distributed shorter series when data of another associated normally distributed longer series is utilized in a composite manner. At the same time, it may be seen if the longer series also gains from this approach. As all required variance-covariance matrices in this case are available in closed form [*Rueda*, 1981], the relative information for parameters of longer and shorter series can be obtained analytically. Considering N_X and N_Y observations of X and Y, the asymptotic variance-covariance matrix of parameter estimates are obtainable in closed form by matrix inversion of equation (20) as

$$VC_X\left(\mu_X, \sigma_X^2\right) = \frac{1}{N_X}\begin{Vmatrix} \sigma_X^2 & 0 \\ & 2\left(\sigma_Y^2\right)^2 \end{Vmatrix} \tag{25}$$

and

$$VC_Y\left(\mu_Y, \sigma_Y^2\right) = \frac{1}{N_Y}\begin{Vmatrix} \sigma_Y^2 & 0 \\ & 2\left(\sigma_Y^2\right)^2 \end{Vmatrix}. \tag{26}$$

Using equation (20) and (24), the elements of the Fisher information matrix for the simplified composite event in Figure 1, i.e., with $m_Y = 1$, can be obtained from equation (12) as

$$I_C\left(\mu_X, \sigma_X^2, \mu_Y, \sigma_y^2, \theta\right) = n_{XY}\begin{Vmatrix} \frac{1}{\sigma_X^2}\left[(m_X - 1) + \frac{1}{(1-\theta^2)}\right] & 0 & \frac{-\theta}{\sigma_X\sigma_Y(1-\theta^2)} & 0 & 0 \\ & \frac{1}{2(\sigma_X^2)^2}\left[(m_X - 1) + \frac{(2-\theta^2)}{2(1-\theta^2)}\right] & 0 & \frac{-\theta^2}{4\sigma_X^2\sigma_Y^2(1-\theta^2)} & \frac{-\theta}{2\sigma_X^2(1-\theta^2)} \\ & & \frac{1}{\sigma_Y^2(1-\theta^2)} & 0 & 0 \\ & & & \frac{(2-\theta^2)}{4(\sigma_Y^2)^2(1-\theta^2)} & \frac{-\theta}{2\sigma_Y^2(1-\theta^2)} \\ & & & & \frac{1+\theta^2}{(1-\theta^2)^2} \end{Vmatrix}.$$

Table 7. Percent Information Gain for Distribution Parameter Estimates for Shorter Series Y

	Marginals		% Information Gain for Parameter Estimates of Y at $m_x = 3$							
			For Moderate ρ				For High ρ			
Case	X	Y	ρ	$\hat{\mu}_Y/\hat{\gamma}_Y$	$\hat{\sigma}_Y^2/\hat{\alpha}_Y$	$\hat{\beta}_Y$	ρ	$\hat{\mu}_Y/\hat{\gamma}_Y$	$\hat{\sigma}_Y^2/\hat{\alpha}_Y$	$\hat{\beta}_Y$
I	NOR	**NOR**	0.7	49	19	–	0.9	117	78	–
II	LEV	**LEV**	0.6	27	33	–	0.67	33	49	–
III	NOR	**LEV**	0.7	76	27	–	0.9	221	137	–
IV	NOR	**G2**	0.7	–	13	14	0.9	–	79	80
V	G2	**LEV**	0.7	76	27	–	0.9	180	92	–
VI	LEV	**LP3**	0.4	8	10	10	0.5	27	35	33

The corresponding variance-covariance matrix can be obtained in closed-form by inverting the above information matrix as

$$VC_C\left(\mu_X,\sigma_X^2,\mu_Y,\sigma_y^2,\theta\right)=\frac{1}{n_{XY}}\left\|\begin{array}{ccccc}\frac{\sigma_X^2}{m_X} & 0 & \frac{\theta\sigma_X\sigma_Y}{m_X} & 0 & 0\\ & \frac{2\left(\sigma_X^2\right)^2}{m_X} & 0 & \frac{2\theta^2\sigma_X^2\sigma_Y^2}{m_X} & \frac{\sigma_X^2\theta(1-\theta^2)}{m_X}\\ & & \frac{\sigma_Y^2[1+(m_X-1)(1-\theta^2)]}{m_X} & 0 & 0\\ & & & \frac{2\left(\sigma_Y^2\right)^2[\theta^4+m_X(1-\theta^4)]}{m_X} & \frac{\sigma_Y^2\theta(1-\theta^2)[\theta^2+m_X(1-\theta^2)]}{m_X}\\ & & & & \frac{(1-\theta^2)^2[\theta^2+m_X(2-\theta^2)]}{2m_X}\end{array}\right\|. \quad (27)$$

The relative information for mean and variance can now be obtained for longer and shorter series using equations (18) and (19) on the basis of variance-covariance matrices given in equations (25), (26), and (27) as

$$RI_X(\mu_X)=\frac{1}{m_X}\left(\frac{b_X^{11}}{b_C^{11}}\right)=\frac{1}{m_X}\frac{\sigma_X^2}{\frac{\sigma_X^2}{m_X}}=1,$$

$$RI_X\left(\sigma_X^2\right)=\frac{1}{m_X}\left(\frac{b_X^{22}}{b_C^{22}}\right)=\frac{1}{m_X}\frac{2\left(\sigma_X^2\right)^2}{\frac{2\left(\sigma_X^2\right)^2}{m_X}}=1,$$

and

$$RI_Y(\mu_Y)=\frac{1}{m_Y}\left(\frac{b_Y^{11}}{b_C^{33}}\right)=\frac{1}{1}\frac{\sigma_Y^2}{\frac{\sigma_Y^2[1+(m_X-1)(1-\theta^2)]}{m_X}}=\frac{1}{1-\theta^2(1-1/m_X)},$$

$$RI_Y\left(\sigma_y^2\right)=\frac{1}{m_Y}\left(\frac{b_Y^{22}}{b_C^{44}}\right)=\frac{1}{1}\frac{2\left(\sigma_Y^2\right)^2}{\frac{2\left(\sigma_Y^2\right)^2[\theta^4+m_X(1-\theta^4)]}{m_X}}=\frac{1}{1-\theta^4(1-1/m_X)}.$$

It is apparent from the above that for the longer series X the relative information remains unity for both mean and variance for any m_X and any association level. In other words, the longer series does not gain at all from the shorter series when both marginals are normal, irrespective of the level of cross correlation. For the shorter series Y, it may be seen that, for $m_X = 1$, the relative information for both parameters is unity, irrespective of the association level. Furthermore, when the two series are independent, i.e., when $\theta = 0$, relative information is again unity for both parameters, irrespective of the value of length ratio m_X. However, when there is a perfect association, i.e., when $\theta = 1$, the relative information for both parameters is equal to m_X. In other words, the relative information curve is a straight line passing through origin and having unit slope and thus for $m_X = 3$, for example, both these gains would be 200%.

[43] The graphical and tabular results of the expected information gain for mean and variance parameters of the shorter series, $RI_{pp}\ \forall\ p = \{3,4\} = \{\mu_Y,\sigma_Y^2\}$, are given in Figure 7 and Table 7, respectively. It may be seen from these plots and table that there is a greater gain in information for mean as compared to variance. For a value of $m_X = 3$ and $\rho = 0.7$, the gain in mean is 49%, and for variance, it is 19%. For a higher correlation of $\rho = 0.9$, the gains for both increase rather sharply and are 117% and 78%, respectively.

4.2.2. Case II: LEV Marginals

[44] In this non-Gaussian case, the objective is to see up to what degree both the shorter and longer Gumbel distributed series benefit each other in terms of relative information. The relative information for the longer and shorter series is obtained using equations (18) and (19) on the basis of information matrices for univariate Gumbel distribution given in equation (21) and that obtained numerically for the composite likelihood approach involving the bivariate Gumbel Type A distribution. The graphical and tabular results for relative information for location and scale parameters of longer and

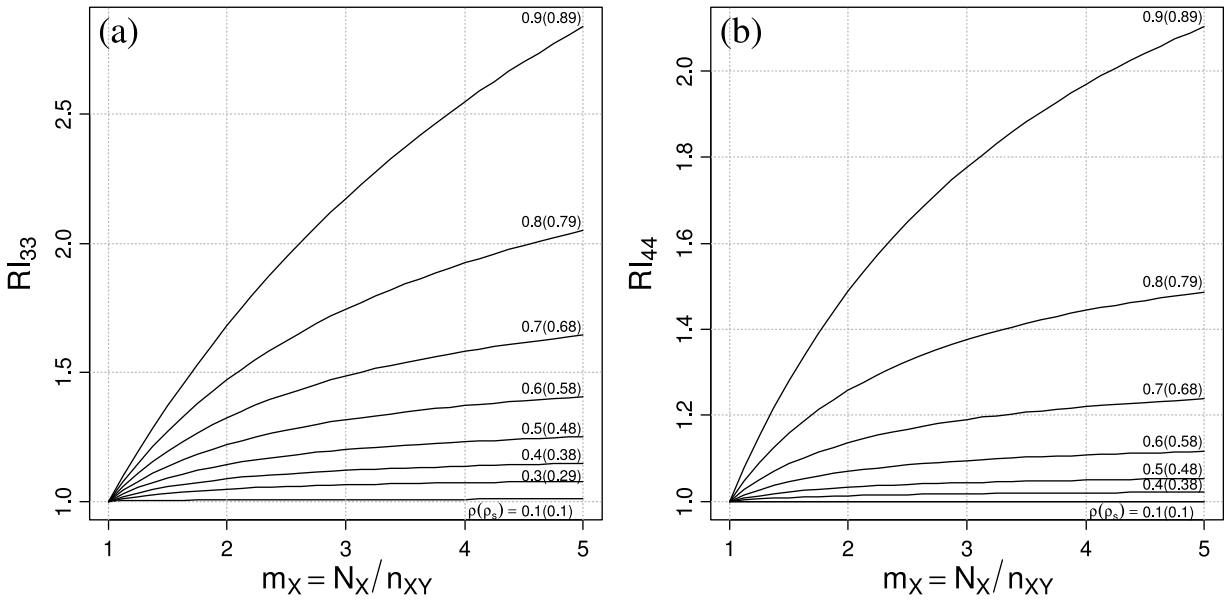

Figure 7. Relative information RI_{pp} for parameter estimates of an incomplete bivariate normal data as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to mean and variance parameters $\{3, 4\} \equiv \{\mu_Y, \sigma_Y^2\}$ of shorter series *Y*.

shorter series, $RI_{pp} \forall p = \{1,2,3,4\} = \{\gamma_X,\alpha_X,\gamma_Y,\alpha_Y\}$, are given in Figure 8 and Tables 6 and 7, respectively. It may be seen that, for $\rho = 0.6$, the gain for location parameter γ_X of longer series is insignificant at the 2% level. For the scale parameter though, the gain is higher at 12%. For a higher $\rho = 2/3$, which is the maximum admissible value of correlation for this distribution, these gains are 3% and 29%, respectively. The shorter-series gains are significant for both parameters, gains being 27% and 33% for $\rho = 0.6$ and 33% and 49% for $\rho = 2/3$, respectively.

4.2.3. Case III: NOR and LEV Marginals

[45] This case is a combination of Gaussian and non-Gaussian distributions, with the Frank copula providing the basis for their joint distribution. It would be important to see if the relative information characteristics in this case are significantly different from the previous two cases. The relative information is obtained for longer and shorter series using equations (18) and (19) on the basis of information matrices for normal and largest extreme value distributions given in equation (20) and (21) and that obtained numerically for the Frank copula-based bivariate distribution. The graphical and tabular results for mean and variance and location and scale parameters of longer and shorter series, $RI_{pp} \forall p = \{1,2,3,4\} = \{\mu_X,\sigma_X^2,\gamma_Y,\alpha_Y\}$, are given in Figure 9 and Tables 6 and 7, respectively. For $\rho = 0.7$, the gains for mean and variance of longer series are 18% and 8%, respectively. For a higher correlation of $\rho = 0.9$, these gains are 65% and 41%, respectively. These gains have interesting comparison with Case I wherein there were no such gains for normally distributed longer series. The shorter-series gains are significantly greater for both location and scale parameters at 76% and 27% for $\rho = 0.7$ and at 221% and 137% for $\rho = 0.9$, respectively. Comparison with corresponding gains in Case II, having LEV distributed longer series, indicates that these gains are substantially higher.

4.2.4. Case IV: NOR and G2 Marginals

[46] This case is similar to Case III, except for the gamma distribution replacing the largest extreme value distribution for the shorter series. It would be important to see if the relative information characteristics for normally distributed longer series are different in the two cases. The relative information is obtained for the two series using equations (18) and (19), on the basis of information matrices for normal and gamma distributions given in equations (20) and (22), and that obtained numerically for the composite likelihood approach involving the Frank copula-based bivariate distribution. The graphical and tabular results of the relative information for mean and variance and scale and shape parameters of longer and shorter series, $RI_{pp} \forall p = \{1,2,3,4\} = \{\mu_X,\sigma_X^2,\alpha_Y,\beta_Y\}$, are given in Figure 10 and Tables 6 and 7, respectively. It may be seen that, for $\rho = 0.7$, the gains for mean and variance of longer series are 16% and 6%, respectively. For a higher $\rho = 0.9$, these gains are 51% and 24%, respectively. All these gains are less than those in the previous case of shorter series being largest extreme value distributed. However, this result cannot be generalized, as it pertains to the specific shape parameter $\beta_Y = 12$. The gains for the shorter series for both scale and shape parameters are moderate at 13% and 14%, respectively. For $\rho = 0.9$ though, the gains are significantly higher at 79% and 80%, respectively.

4.2.5. Case V: G2 and LEV Marginals

[47] This case is similar to Case III, except for the longer series being gamma distributed instead of normal distributed. It would be important to see if the relative information characteristics for the shorter largest extreme value distributed series are different in the two cases. The relative information is obtained for the two series using equations (18) and (19) on the basis of information matrices for gamma and largest extreme value distributions given in equations (22)

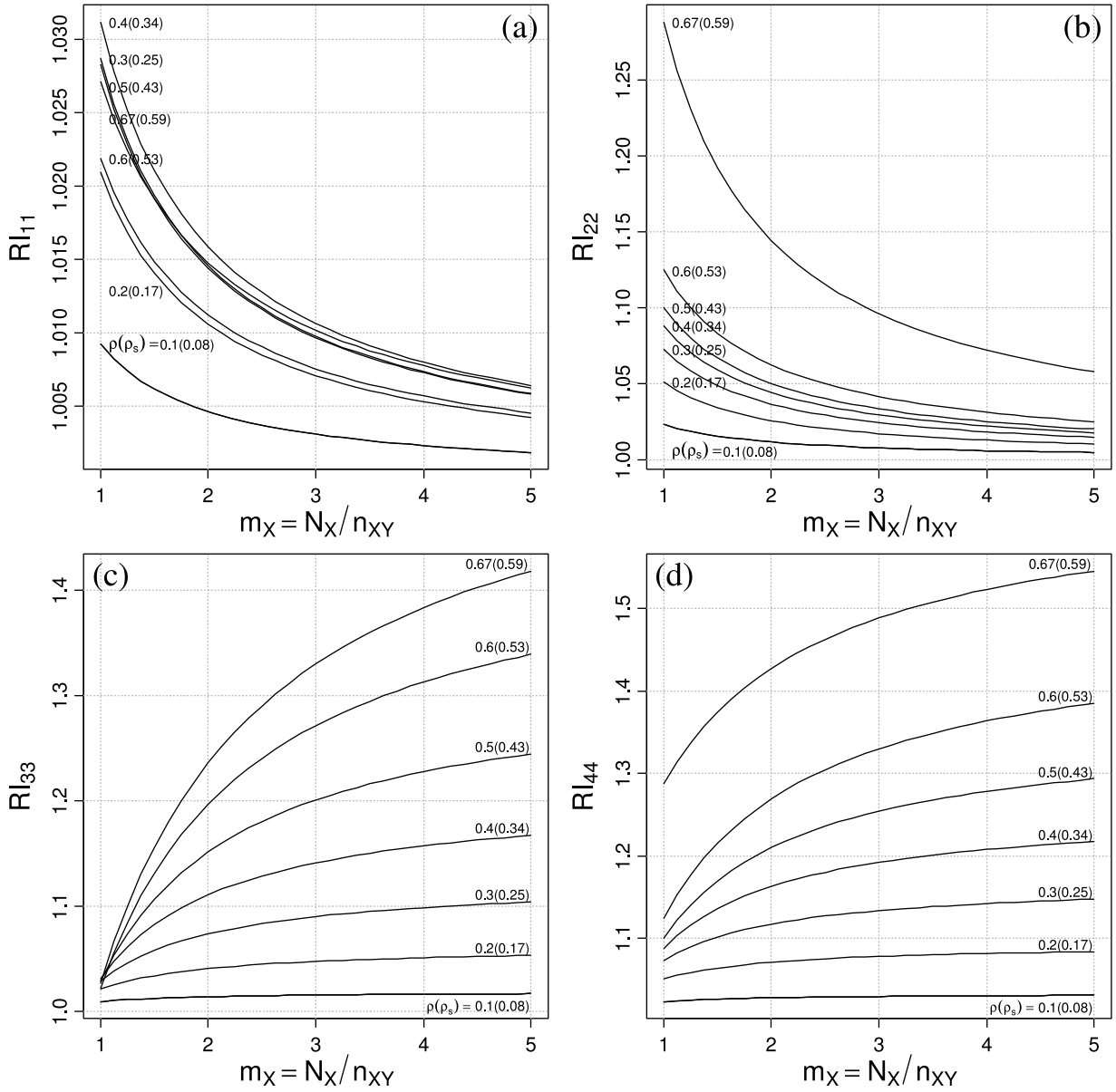

Figure 8. Relative information RI_{pp} for parameter estimates of an incomplete bivariate Gumbel Type A data as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to location and scale parameters $\{1:2\} \equiv \{\gamma_X, \alpha_X\}$ of longer series X and $\{3:4\} \equiv \{\gamma_Y, \alpha_Y\}$ of shorter series Y.

and (21) and that obtained numerically for the composite likelihood approach involving the Frank copula-based bivariate distribution. The graphical and tabular results for relative information for scale and shape parameters and location and scale parameters of longer and shorter series, $RI_{pp}\ \forall\ p = \{1,2,3,4\} = \{\alpha_X,\beta_X,\gamma_Y,\alpha_Y\}$, are given in Figure 11 and Tables 6 and 7, respectively. For a moderate $\rho = 0.7$, gains for scale and shape parameters of longer series are marginal at 8% each. For a higher $\rho = 0.9$, these gains are moderately higher at 30% and 29%, respectively. For moderate $\rho = 0.7$, the gains for shorter series are higher for location than for scale and are 76% and 27%, respectively. In fact, the levels of these gains are same as that in Case III. However, a generalized statement to this effect may not be appropriate in view of the specific shape parameter of $\beta_Y = 12$. For a higher $\rho = 0.9$, the gains are significantly greater at 180% and 92%, respectively. However, these gains are substantially less than those in Case III.

4.2.6. Case VI: LEV and LP3 Marginals

[48] This case involves an LP3 distributed marginal and may have direct relevance for flood frequency analysis in the United States where it is the prescribed distribution. The relative information is obtained for the two series using equations (18) and (19) on the basis of information matrices for LEV and LP3 distributions given in equations (21) and (23) and that obtained numerically for the composite likelihood approach involving the Frank copula-based bivariate distribution. Graphical and tabular results for the relative information for location and scale parameters and location, scale, and shape parameters of longer and shorter series,

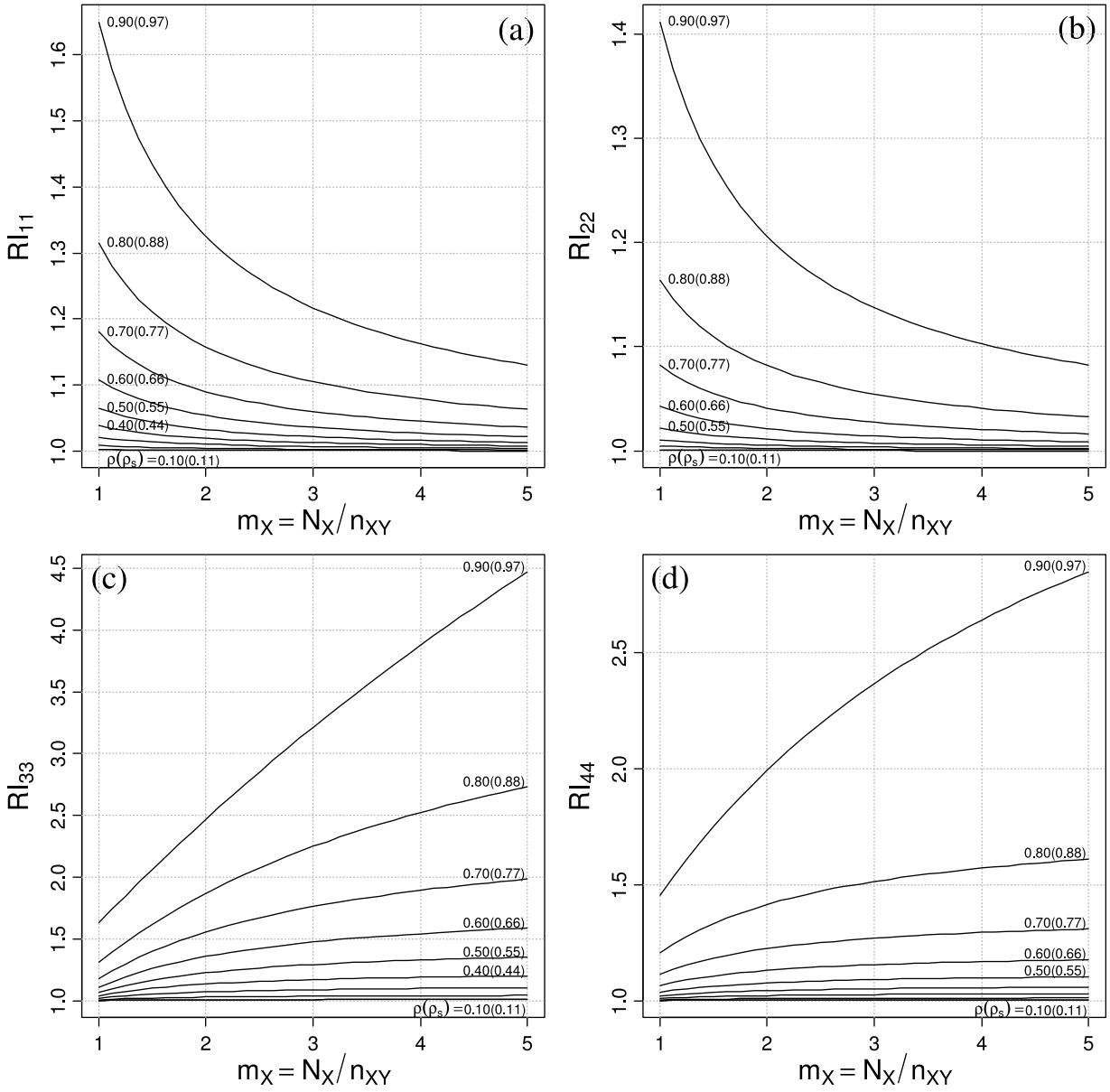

Figure 9. Relative information RI_{pp} for parameter estimates of an incomplete Frank copula-based bivariate data, with NOR and LEV marginals, as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to the mean and variance parameters $\{1:2\} \equiv \{\mu_X, \sigma_X^2\}$ of longer series X and the location and scale parameters $\{3:4\} \equiv \{\gamma_Y, \alpha_Y\}$ of shorter series Y.

$RI_{pp} \forall p = \{1,2,3,4\} = \{\gamma_X, \alpha_X, \gamma_Y, \alpha_Y, \beta_Y\}$, are given in Figure 12 and Tables 6 and 7. The linear correlation coefficient ρ hugely underestimates the actual association in this case that may be of nonlinear nature. For example, values of $\rho = 0.4$ and $\rho = 0.5$ correspond to Spearman correlation coefficients of $\rho_s = 0.78$ and $\rho_s = 0.92$, respectively, indicating higher association. These two values will therefore be used in this case while referring to moderate and higher association levels. It may be seen that, for $\rho = 0.4$, gains for location and scale parameters of the longer series are moderate at 18% and 12%, respectively. For a higher $\rho = 0.5$, these gains increase to 40% and 27%, respectively. Gains for the shorter series are marginal for all three parameters at 8%, 10%, and 10%, respectively. For the higher correlation of $\rho = 0.5$, gains are comparatively higher at 27%, 35%, and 33%, respectively. Another observation in this case is that gains for LP3 series do not grow significantly for increasing values of m_X.

5. Discussion

[49] The theoretical basis of the proposed composite likelihood approach is elaborated in section 2, wherein equations (18) and (19) provide expressions for information gain through simultaneous consideration of concurrent and nonconcurrent portions of incomplete bivariate data sets. The first application in section 3 demonstrates the advantage of the approach wherein flood quantiles for 100, 200, and 500 year return periods are obtained with about 30% less standard error or, in other words, with 30% better confidence, as compared to the conventional univariate analysis.

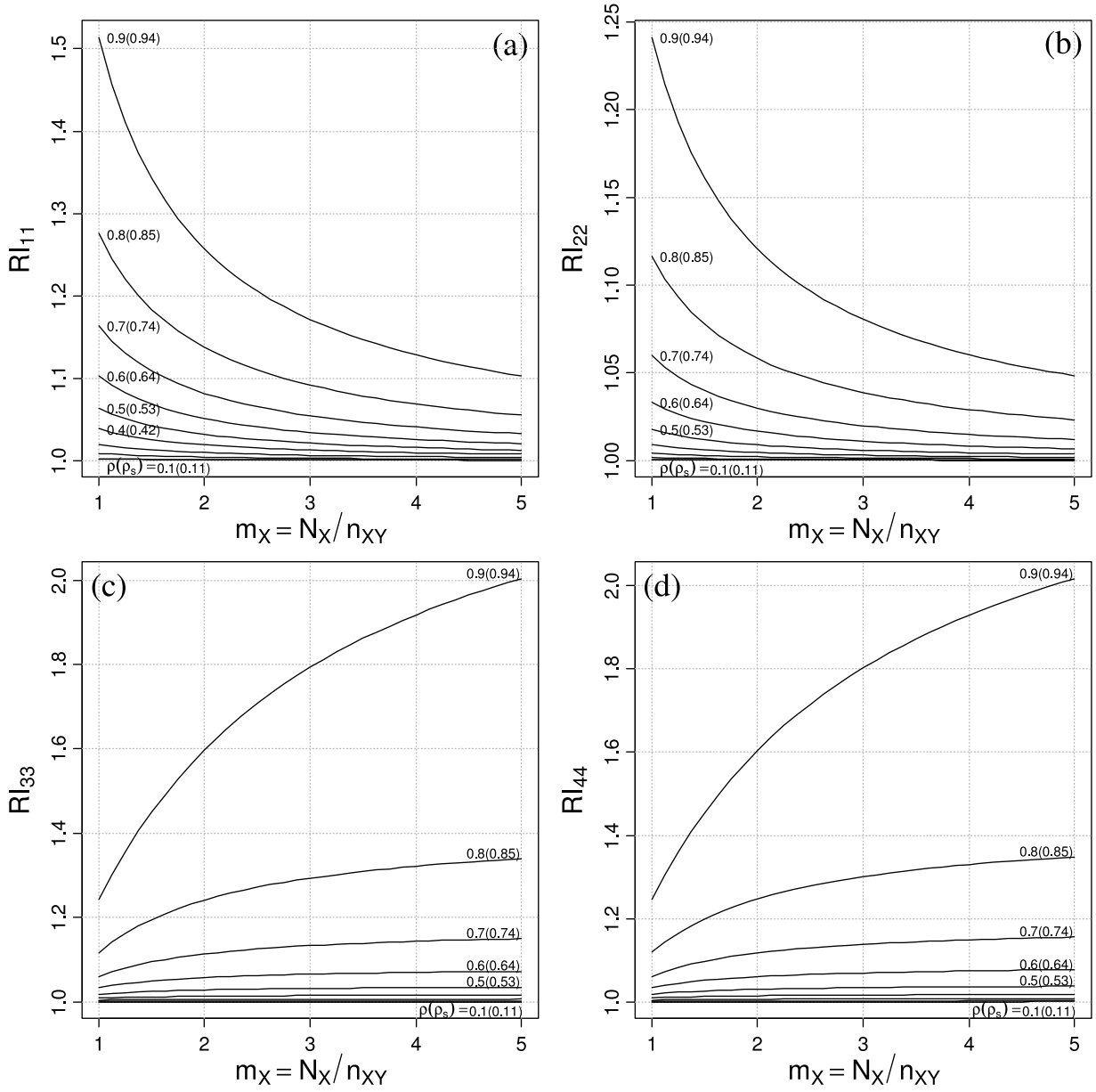

Figure 10. Relative information RI_{pp} for parameter estimates of an incomplete Frank copula-based bivariate data, with NOR-G2 marginals, as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to the mean and variance parameters $\{1:2\} \equiv \{\mu_X, \sigma_X^2\}$ of longer series X and the scale and shape parameters $\{3:4\} \equiv \{\alpha_Y, \beta_Y\}$ of shorter series Y.

This improvement pertained to the case where shorter annual peak flow data series benefited from an associated downstream flow series having 40% more data (i.e., having the length ratio $m_X = 1.4$). Reduction in uncertainty can be as high as 50%–60% for $m_X = 2.0$, which is not uncommon in an insufficient data availability situation. The second application in section 4, involving six different combinations of Gaussian and non-Gaussian marginals, establishes accrual of significant "expected information gain" through this approach. For these six cases, the average and maximum expected gains for moderate level of association are about 30% and 75%, respectively. For higher association levels, these gains are as high as 90% and 220%, respectively. Although, these six cases pertained to specific marginals and the Frank copula, the results are indicative of the impressive gains that are achievable at no extra cost for other combinations of marginals and copula models.

[50] Every hydrologic design estimate has associated uncertainty, and keeping this uncertainty at an acceptable level continues to be an important design objective. Desirable reduction in uncertainty in hydrologic design estimates that are predominantly based on univariate frequency analysis is typically achieved by operating observation networks for longer periods and therefore have direct cost consequences. The proposed approach marks a paradigm shift in the strategy for reducing uncertainty by providing an alternative avenue for increasing information by pooling staggered but associated hydrologic data that already exist and thereby avoiding the need for extra funding to run networks for extended periods. Despite a greater need for

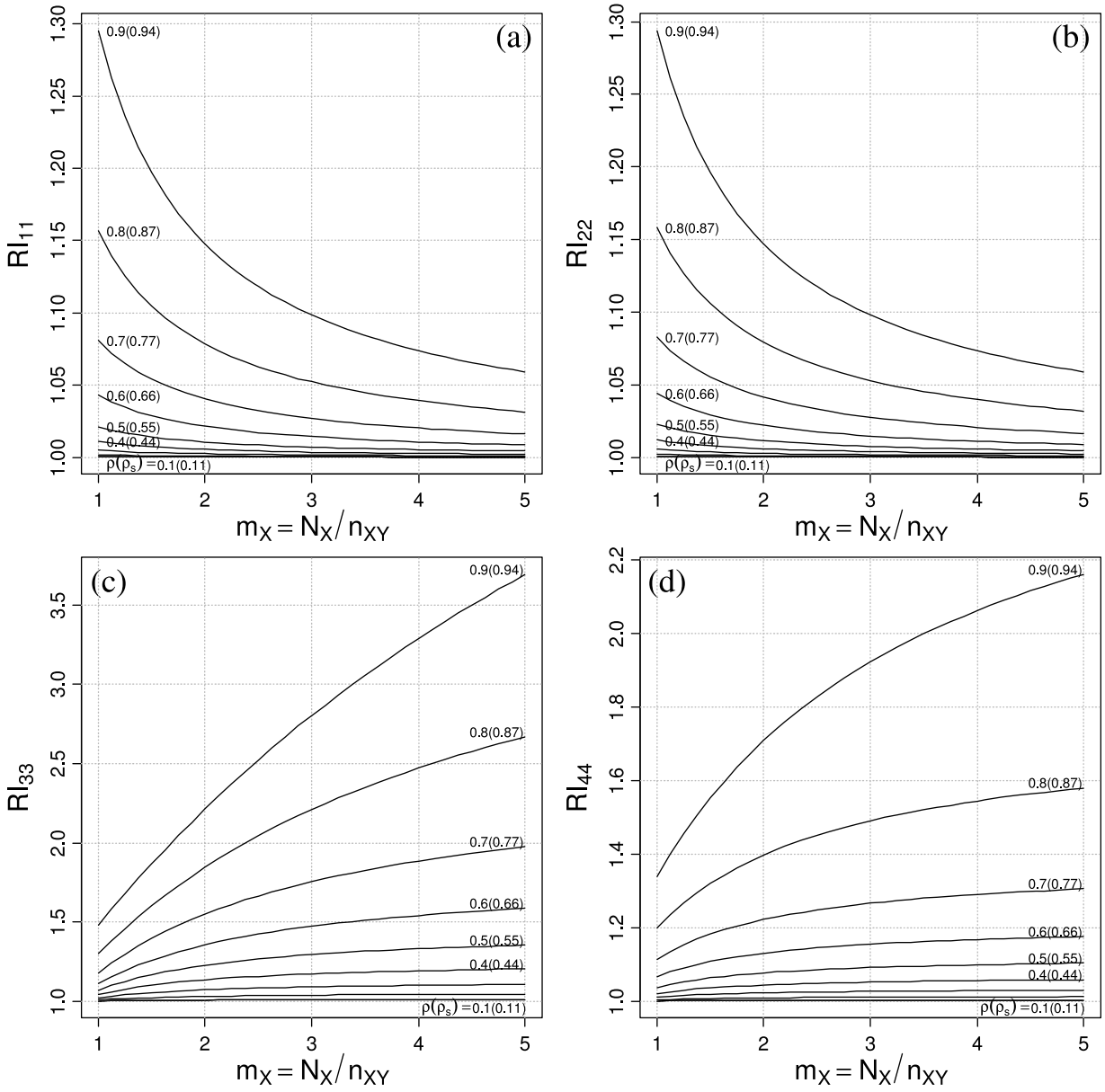

Figure 11. Relative information RI_{pp} for parameter estimates of an incomplete Frank copula-based bivariate data, with G2-LEV marginals, as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to the scale and shape parameters $\{1 : 2\} \equiv \{\alpha_X, \beta_X\}$ of longer series X and the location and scale parameters $\{3 : 4\} \equiv \{\gamma_Y, \alpha_Y\}$ of shorter series Y.

strengthening hydrologic observation networks in view of tremendous pressure on water as a resource and for studying potential impacts of climate change, hydrologic observation networks in many countries have been dwindling in recent times [*IAHS*, 2001], mainly due to financial constraints and lack of awareness at governmental levels. As information for the discontinued stations or data series can be enhanced by employing the proposed approach, it can significantly offset the negative impact of shrinking networks. At the same time, the methodology can also be advantageously applied for enhancing precision pertaining to shorter data lengths at newer stations in an upgraded network. Hydrological data sets typically abound varied length data series, as different stations or variables have different and staggered observation periods. The approach has a high practical value, as it is particularly suited for shorter hydrologic data series that are often considered inadequate for yielding satisfactory design estimates. In such situations of data inadequacy, this approach can help enhance precision by integrating longer data series of the same or other associated processes that are invariably available from the existing baseline or other long-term observation stations in the vicinity. A variety of hydrological applications involving rainfall, flow, soil moisture, groundwater, and/or water quality processes can benefit from this approach, wherein additional information for shorter data series is gained.

6. Conclusions

[51] Increasingly better parameter estimation methods for hydrological frequency distributions that yield parameters

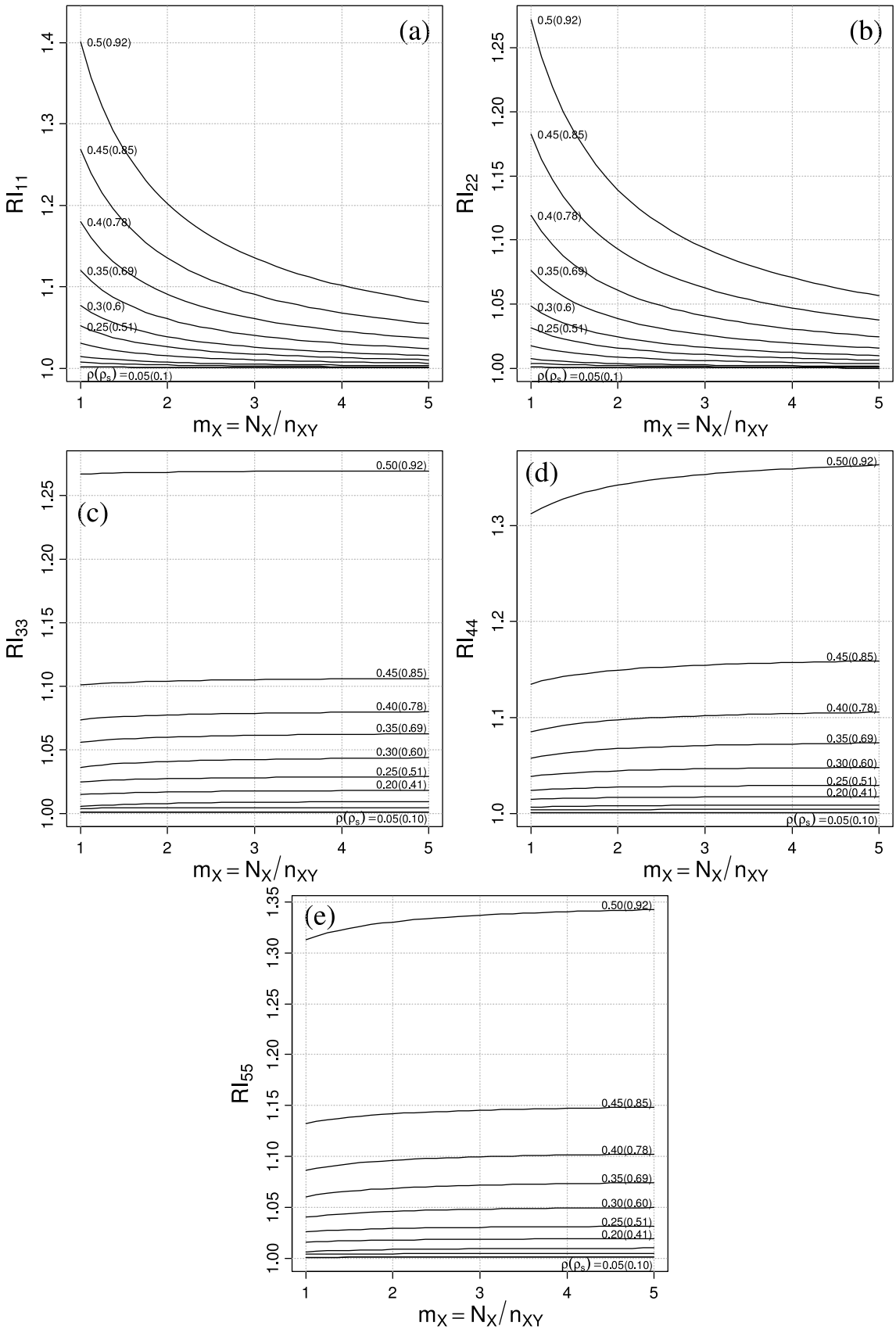

Figure 12. Relative information RI_{pp} for parameter estimates of an incomplete Frank copula-based bivariate data, with LEV-LP3 marginals, as a function of length ratio m_X and association levels ρ (or ρ_s). Subscripts "*pp*" correspond to the location and scale parameters $\{1 : 2\} \equiv \{\gamma_X, \alpha_X\}$ of longer series X and the location, scale, and shape parameters $\{3 : 5\} \equiv \{\gamma_Y, \alpha_Y, \beta_Y\}$ of shorter series Y.

with minimum variance or higher precision have evolved as a result of considerable research effort of statisticians and hydrologists. The composite likelihood approach presented in this paper gives a new dimension to this evolution by offering a mechanism, wherein arbitrarily distributed longer and shorter data series can be integrated in a copula-based multivariate framework, yielding parameter estimates with higher precision. Specific results of impressive information gains in two applications included in this paper have been discussed. A few general and some specific conclusions arising from these applications are as follows:

[52] 1. Significant information gain is achievable for shorter series even at moderate association levels.

[53] 2. There is information gain even for the longer series, except for the bivariate normal case.

[54] 3. Gains for longer series decrease with increasing values of the length ratio m_X, whereas they increase for shorter series. For a given length ratio m_X, gains are proportional to the level of association.

[55] 4. For shorter series in a bivariate normally distributed data, the limiting gains are equal to the length ratio m_X when there is near perfect association. This indicates a tremendous benefit of this approach when two highly correlated normal data series are under consideration.

[56] 5. Gains for location parameter for shorter largest extreme value series are substantially higher when combined with a longer normally distributed series, as compared to that when combined with another largest extreme value series. These gains are exactly the same when the longer series is gamma distributed.

[57] Overall, the proposed composite likelihood approach offers a promising framework for integrating staggered information that has hitherto remained unharnessed for obtaining hydrologic design estimates. The approach yields more information from the existing baseline and other long-term observation stations at no extra cost and thus can play a vital role in offsetting losses due to declining data availability in the current tough economic times. The approach is of particular advantage for designs pertaining to partially ungauged basins or for recently upgraded networks in which many data series have shorter record lengths. Further studies may be aimed at characterizing potential information gain for other marginals and copula models pertinent to different hydrologic designs and at evaluating benefits of considering higher order multivariate models. In view of its simplicity and character of significantly improving the precision, this approach has a potential for faster adoption by researchers and field practitioners alike.

[58] **Acknowledgments.** The authors thank the editors and four anonymous referees for the helpful remarks and comments, which significantly improved the quality of the paper. This work is based on the first author's Ph.D. dissertation research, and the partial financial support received by him from the Graduate School of Louisiana State University to pursue the same is most sincerely acknowledged.

References

Abberger, K. (2005), A simple graphical method to explore tail-dependence in stock-return pairs, *Appl. Financ. Econ.*, *15*(1), 43–51.

Anderson, T. W. (1957), Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *J. Am. Stat. Assoc.*, *52*(278), 200–203.

Buishand, T. A. (1984), Bivariate extreme-value data and the station-year method, *J. Hydrol.*, *69*, 77–95.

Chowdhary, H. (2010), Copula-based multivariate hydrologic frequency analysis, Ph.D. thesis, Louisiana State Univ., Baton Rouge, La.

Clarke, R. T. (1980), Bivariate gamma distributions for extending annual streamflow records from precipitation: Some large-sample results, *Water Resour. Res.*, *16*(5), 863–870, doi:10.1029/WR016i005p00863.

Cox, D. R., and D. V. Hinkley (1974), *Theoretical Statistics*, Chapman and Hall, London.

Dalrymple, T. (1960), Flood frequency analysis, *U.S. Geol. Surv., Water Suppl. Pap.*, *1543-A*.

De Michele, C., G. Salvadori, M. Canossi, A. Petaccia, and R. Rosso (2005), Bivariate statistical approach to check adequacy of dam spillway, *J. Hydrol. Eng.*, *10*(1), 50–57.

Edgett, G. L. (1956), Multiple regression with missing observations among the independent variables, *J. Am. Stat. Assoc.*, *51*, 122–131.

Escalante, C. (2007), Application of bivariate extreme value distribution to flood frequency analysis: A case study of Northwestern Mexico, *Nat. Hazards*, *42*, 37–46.

Escalante, C., and J. A. Raynal-Villasenor (1998), Multivariate estimation of floods: The trivariate gumbel distribution, *J. Stat. Comput. Simul.*, *61*(4), 313–340.

Escalante, C., and J. A. Raynal-Villasenor (2008), Trivariate generalized extreme value distribution in flood frequency analysis, *J. Hydrolog. Sci.*, *53*(3), 550–567.

Favre, A.-C., S. El Adlouni, L. Perreault, N. Thiemonge, and B. Bobee (2004), Multivariate hydrological frequency analysis using copulas, *Water Resour. Res.*, *40*, 1–12, W01101, doi:10.1029/2003WR002456.

Fermanian, J.-D. (2005), Goodness-of-fit tests for copulas, *J. Multivar. Anal.*, *95*, 119–152.

Fiering, M. B. (1962), On the use of correlation to augment data, *J. Am. Stat. Assoc.*, *57*(297), 20–32.

Fisher, N. I., and P. Switzer (2001), Statistical computing and graphics, *J. Am. Stat.*, *55*(3), 233–239.

Genest, C., and J.-C. Boies (2003), Detecting dependence with Kendall plots, *J. Am. Stat.*, *57*(4), 275–284.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrol. Eng.*, *12*(4), 347–368.

Genest, C., J.-F. Quessy, and B. Remillard (2006), Goodness-of-fit procedures for copula models based on the probability integral transformation, *Scand. J. Stat.*, *33*(2), 337–366.

Grimaldi, S., F. Serinaldi, F. Napolitano, and L. Ubertini (2005), A 3-copula function application for design hyetograph analysis, *IAHS-AISH Publ.*, *293*, 203–211.

Gumbel, E. J., and C. K. Mustafi (1967), Some analytical properties of bivariate extremal distributions, *J. Am. Stat. Assoc.*, *62*, 569–588.

IAHS (2001), Ad Hoc Group on Global Water Data Sets. Global water data: A newly endangered species, *Eos Trans. AGU*, *82*(5), 54, 56, 58.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, Boca Raton, Fla.

Kao, S., and R. S. Govindaraju (2007a), Probabilistic structure of storm surface runoff considering the dependence between average intensity and storm duration of rainfall events, *Water Resour. Res.*, *43*, W06410, doi:10.1029/2006WR005564.

Kao, S., and R. S. Govindaraju (2007b). A bivariate frequency analysis of extreme rainfall with implications for design, *J. Geophys. Res.*, *112*, D13119, doi:10.1029/2007JD008522.

Kao, S. C., and R. S. Govindaraju (2008), Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas, *Water Resour. Res.*, *44*, W02415, doi:10.1029/2007WR006261.

Lord, F. M. (1955), Estimation of parameters from incomplete data, *J. Am. Stat. Assoc.*, *50*, 870–876.

Matalas, N. C., and B. Jacobs (1964), A correlation procedure for augmenting hydrologic data, *U.S. Geol. Surv. Prof. Pap.*, *434-E*.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974), *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer, New York.

Poulin, A., D. Huard, A.-C. Favre, and S. Pugin (2007), Importance of tail dependence, *J. Hydrol. Eng.*, *12*(4), 394–403.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, Wiley, New York.

Raynal-Villasenor, J. A. (1985), Bivariate extreme value distributions applied to flood frequency analysis, Ph.D. thesis, Colorado State Univ., Fort Collins, Colo.

Raynal-Villasenor, J. A., and J. D. Salas (2008), Using bivariate distributions for flood frequency analysis based on incomplete data, *Proc. World Environ. Water Resour. Cong.*, *ASCE*, Honolulu, Hawaii.

Rueda, E. (1981), Transfer of information for flood related variables, M.S. thesis, Colorado State Univ., Fort Collins, Colo.

Salvadori, G., and C. De Michele (2004), Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, *Water Resour. Res.*, *40*, W12511, doi:10.1029/2004WR003133.

Salvadori, G., and C. De Michele (2007), On the use of copulas in hydrology: Theory and practice, *J. Hydrol. Eng.*, *12*(4), 369–380.

Samaniego, L., L. A. Bardossy, and R. Kumar (2010), Streamflow prediction in ungauged catchments using copula-based dissimilarity measures, *Water Resour. Res.*, *46*, W02506, doi:10.1029/2008WR007695.

Serinaldi, F. (2009), A multisite daily rainfall generator driven by bivariate copula-based mixed distributions, *J. Geophys. Res.*, *114*, D10103, doi:10.1029/2008JD011258.

Serinaldi, F., and S. Grimaldi (2007), Fully nested 3-copula: Procedure and application on hydrologic data, *J. Hydrol. Eng.*, *12*(4), 420–430.

Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared, *Water Resour. Res.*, *21*(9), 1421–1432, doi:10.1029/WR021i009p01421.

Vandenberghe, S., N. E. C. Verhoest, and B. De Baets (2010), Fitting bivariate copulas to the dependence structure between storm characteristics: A detailed analysis based on 105 year 10 min rainfall, *Water Resour. Res.*, *46*, W01512, doi:10.1029/2009WR007857.

Wang, W., and M. T. Wells (2000), Model selection and semiparametric inference for bivariate failure-time data, *J. Am. Stat. Assoc.*, *95*(449), 62–72.

Wilks, S. S. (1962), *Mathematical Statistics*, Wiley, New York.

Zhang, L., and V. P. Singh (2006), Bivariate flood frequency analysis using the copula method, *J. Hydrol. Eng.*, *11*(2), 150–164.

Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *J. Hydrol.*, *332*, 93–109.

H. Chowdhary, Department of Civil and Environmental Engineering, Louisiana State University, Baton Rouge, LA 70803, USA. (hchowd1@lsu.edu)

V. P. Singh, Department of Biological and Agricultural Engineering and Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843, USA. (vsingh@tamu.edu)